6/

REPORT RESUMES

ED 003 255 24
ITEM ANALYSIS, TEST DESIGN, AND CLASSIFICATION.
BY- SOLOMON, HERBERT
STANFORD UNIV., CALIF.
REPORT NUMBER CRP-1327 PUB DATE 65
EDRS PRICE MF-$0.54 HC-$15.40 385P.

DESCRIPTORS- MATHEMATICS, *MATHEMATICAL APPLICATIONS, MATHEMATICAL CONCEPTS, *MODELS, PROBABILITY, *MEASUREMENT INSTRUMENTS, RESEARCH METHODOLOGY, *DATA ANALYSIS, *TEST CONSTRUCTION, TEST SELECTION, TESTS, QUESTIONNAIRES, STATISTICAL ANALYSIS, FACTOR ANALYSIS, STANFORD, CALIFORNIA

THE THEME OF THIS PROGRAM WAS THE USE OF PROBABILISTIC MODELS AS MEASUREMENT TOOLS FOR EXPERIMENTAL DESIGN IN TEST CONSTRUCTION AND FOR ANALYSIS OF TEST DATA. SOME CENTRAL ISSUES IN THIS FIELD ARE DISCUSSED AND RESOLVED IN THE FIRST FOUR CHAPTERS OF THE REPORT UNDER THE FOLLOWING TITLES--(1) CONTRIBUTIONS TO ITEM SELECTION, (2) CLUSTERING PROCEDURES, (3) ASYMPTOTIC EFFICIENCES OF SOME TESTS OF INDEPENDENCE USED ON ITEM SELECTION, AND (4) THE SCORING OF MULTIPLE-CHOICE QUESTIONNAIRES. THE REMAINING CHAPTERS ARE STUDIES OF A NUMBER OF METHODOLOGICAL ISSUES WHICH AROSE IN THE CONSIDERATION OF THE ABOVE CENTRAL ISSUES, AS FOLLOWS--(1) THE USE AND INTERPRETATION OF PRINCIPAL COMPONENT ANALYSIS IN APPLIED RESEARCH, (2) TESTING FOR THE EQUALITY OF CORRELATION COEFFICIENTS FOR VARIOUS MULTIVARIATE MODELS, (3) ASYMPTOTIC DISTRIBUTION OF FUNCTIONS OF A CORRELATION MATRIX, (4) ORDER STATISTICS IN TWO DIMENSIONS, (5) A DECOMPOSITION OF DICHOTOMOUS RANDOM VARIABLES INTO INDEPENDENT COMPONENTS OF VARIATION, AND (6) SOME APPLICATIONS OF FINITE BOOLEAN ALGEBRAS TO DATA ANALYSIS. (LP)

ED 003255

U. S. DEPARTMENT OF HEALTH, EDUCATION AND WELFARE
Office of Education
This document has been reproduced exactly as received from the person or organization originating it. Points of view or opinions stated do not necessarily represent official Office of Education position or policy.

ITEM ANALYSIS, TEST DESIGN, AND CLASSIFICATION

Cooperative Research Project No. 1327

Herbert Solomon
Director

Stanford University
Stanford, California
1965

The research reported herein was supported by the Cooperative Research Program of the Office of Education, U.S. Department of Health, Education, and Welfare.

PRECEDING PAGE BLANK-NOT FILMED

PREFACE

The ten chapters in this final report are the work of a number of scholars who looked into several problems arising in item selection, test design, and clustering procedures. The theme of the program which brought these researchers together is the use of probabilistic models as measurement tools for experimental design in test construction and for analysis of test data. Chapters I, II, III, IV discuss and resolve some central issues. The remaining chapters that follow study a number of methodological issues which arise in their consideration.

An interesting feature presented in Chapters I and II is the use of sampling to achieve results in item selection and in cluster allocation which are in the neighborhood of optimal solutions. To achieve the optimal solution in these problem settings, one is injected quickly into unmanageable or excessive enumerations. The use of probability sampling makes it possible to arrive at solutions which are in the neighborhood of "best solutions" for preassigned risks of error. A resulting trade-off table can aid those who make the decision. This is one of the most important developments arising from the program and will undoubtedly receive more attention by researchers in this field.

A large debt is owed to the following for their contributions to the research program: P.K. Bhattacharya, Herman Chernoff, J.J. Fortier, Ingram Olkin, Juha Partanen, Minoru Siotani, and C.R. Rao. Thanks are due to Mrs. Carolyn Knutsen for so skillfully typing this massive mathematical manuscript.

PRECEDING PAGE BLANK- NOT FILMED

TABLE OF CONTENTS

Page

PREFACE iii

CHAPTER

I Contributions to Item Selection 1

II Clustering Procedures 75

III Asymptotic Efficiencies of Some Tests of Independence Used in Item Selection 98

IV The Scoring of Multiple Choice Questionnaires 137

V The Use and Interpretation of Principal Component Analysis in Applied Research 169

VI Testing for the Equality of Correlation Coefficients for Various Multivariate Models 219

VII Asymptotic Distribution of Functions of a Correlation Matrix 238

VIII Order Statistics in Two-Dimensions 262

VIV A Decomposition of Dichotomous Random Variables Into Independent Components of Variation 321

X On Some Applications of Finite Boolean Algebras to Data Analysis 360

CONTRIBUTIONS TO ITEM SELECTION

by

Jean Jacques Fortier

INTRODUCTION AND SUMMARY

The central situation in item analysis is one where item variables; i.e., variables whose values are responses to items on a test, questionnaire, inventory, or scale are used to predict a criterion variable. This criterion is usually a one dimensional real-valued quantity. In the development in this paper we are interested in the more general situation where the criterion is vector-valued with real-valued components. This situation can arise in many ways. For instance, the non-observable (latent) criterion might itself be measured through a number of indicator items. Thus, in item selection where a subset of items is compared to a criterion the criterion may be the whole set of items. Or we may wish to compare two sets of items, where one set of items is made the criterion. This would be the case if one would compare two traits say, attitude toward science with I.Q.; both measured through a set of items. Symbolically, let us assume we are given two sets of variables $\{X_i\}$ and $\{Y_j\}$ and we want to predict the Y_j's from the X_i's. Then, we may forget for a moment that we are in the context of item analysis and examine the problem in a more abstract way.

Hotelling [1935, 1936] has studied a similar situation and has shown how to reduce symmetrically the two sets of variables under consideration, to what he called "canonical variates." In this paper, we will consider the problem of breaking one set of variables into a number of predictors, considering the other set as unchanged. We will try, in other words, to predict simultaneously each of the Y_j's from functions based on the X_i's without looking for the "best predictable" (as in canonical variates) quantity as a function of the Y_j's.

This statement will receive more explanation as the exposition is developed. If Y is uni-dimensional, there is a frequently used loss function for evaluating prediction, namely the "squared loss" function: $E(Y-\hat{Y})^2$, where $\hat{Y}$ is a function of the X_i's. What we now require, however, is a simultaneous loss function, that is, a function which measures loss through prediction error in each Y_j where each $\hat{Y}_j$ is an appropriate function of the X_i's. If we require an individual estimate $\hat{Y}_j$ or predictor for each component Y_j, it seems natural to minimize:

$$\sum_j \underset{Y_j(X_i)}{E\ E} (\hat{Y}_j - Y_j)^2$$

which henceforth we will write, for simplicity as

$$\sum_j E(\hat{Y}_j - Y_j)^2$$

for which the solution is known to be [See for instance Blackwell and Girshick [1954], or Section IV in this paper]

$$\hat{Y}_j = E(Y_j | X_1, \ldots, X_n)$$

and at this minimum value:

$$\sum_j E(\hat{Y}_j - Y_j)^2 = \sum_j E\,\mathrm{Var}(Y_j | X_1, \ldots, X_n) .$$

On the other hand, suppose we desire a predictor which is the same for all the components Y_j of the vector Y, and satisfies the simultaneous loss function. In our language we label this a simultaneous predictor. In other words we want a function of the X_i's which can be used to predict each and all of the Y_j's at the same time. Thus we want a loss function which takes into account simultaneously the loss when each variable Y_j is predicted by a single $\hat{Y}_o$, which is a function of the X_i's. Therefore we think of minimizing:

$$\sum_j E(\hat{Y}_o - Y_j)^2 .$$

But we will see that at least in the particular case of linear prediction the properties of such an estimate, as a simultaneous predictor are not appealing. For instance, it does not take into account the relations between the Y_j's. In fact, $\hat{Y}_o$ is the linear form whose coefficients for each X_i are the means of the linear regression coefficients for each X_i over all Y_j. This is developed in Section 1. However, it seems that we can allow more flexibility in our estimation procedure without giving up the central idea of a simultaneous predictor if we replace $\hat{Y}_o$, by $b_j \hat{Y}_o$, and minimize.

$$\sum_j E(b_j \hat{Y}_o - Y_j)^2 .$$

The constants b_j which vary with Y_j allow the scale of prediction to vary even if one has a single simultaneous predictor. For example, these constants tell us by how many units each predicted Y_j, namely $b_j \hat{Y}_o$, varies when the predictor itself, namely $\hat{Y}_o$, varies by one unit. In fact $b_j = \dfrac{\mathrm{cov}(\hat{Y}_o - Y_j)}{\mathrm{Var}\, \hat{Y}_o}$. Now suppose, for every given $\hat{Y}_o$ (defined as a real random variable and a function of the X_i's) the b_j's are chosen to minimize $E(b_j\hat{Y}_o - Y_j)^2$. Then because of the quadratic loss function we have adopted (minimum mean square error) and for the set of b_j's which minimize the loss, we may write from linear regression theory

$$E(b_j\hat{Y}_o - Y_j)^2 = \mathrm{Var}(Y_j) - \mathrm{Var}(Y_j)\rho^2_{(b_j\hat{Y}_o)Y_j}$$

$$= \mathrm{Var}(Y_j) - \mathrm{Var}(Y_j)\rho^2_{\hat{Y}_o Y_j}$$

and thus

$$\sum_j E(b_j\hat{Y}_o - Y_j)^2 = \sum_j \mathrm{Var}(Y_j) - \sum_j \mathrm{Var}(Y_j)\rho^2_{\hat{Y}_o Y_j} .$$

The problem is, therefore, to find a $\hat{Y}_o$ as a function of the set $\{X_i\}$ that will maximize

$$\sum_j \mathrm{Var}(Y_j)\rho^2_{\hat{Y}_o Y_j} .$$

It is interesting to view this index in two situations.

(1) When $\hat{Y}_o$ is linear and $\{Y_j\}$ has one component we get the multiple correlation coefficient model.

(2) When Var $Y_j = 1$ and $\{Y_j\} \equiv \{X_i\}$ we have the variance explained by the first principal component in Hotelling's [1933] factor analysis model.

This analysis and this index which is motivated by the linear regression model lead to a solution in estimation or prediction which we have termed simultaneous prediction and is the central development in this study. Suppose we first restrict ourselves to the case of simultaneous linear prediction. Definition: "The best simultaneous linear predictor of a set of real random variables $\{Y_j\}$, $j=1,\ldots,m$, given a set of real random variables $\{X_i\}$, $i=1,\ldots,n$, is that linear combination S of the X's which maximizes the quantity

$$\sum_{j=1}^{m} \mathrm{Var}[Y_j]\rho^2_{Y_jS}$$

where ρ_{Y_jS} is the correlation coefficient between Y_j and S."

One may object at this point that a uni-dimensional predictor is used without justification. But, besides the obvious advantages (ordering purposes for instance) we have satisfied a condition sometimes asked for by sociologists and psychologists, of reducing the dimensionality of the problem. However there should not be a restriction to one dimension and we now present a procedure for constructing step by step a sequence of orthogonal predictors, each removing ("explaining") as much variability from the system as its structure permits. The first Best Simultaneous Linear Predictor, BSLP, accounts for as much variance as its structure permits, then the second BSLP accounts for as much as possible of the variance remaining and so on until all the variation in the model is accounted for. The sequence is constructed step by step by subtracting from the original variables Y_j, the regression on the previously obtained simultaneous predictors, and working on the new variables thus produced. This is developed in Section I.

We therefore have a most economical linear structure in terms of the number of components or dimensions needed. In factor analysis and in classical regression analysis, parsimony through the reduction of number of variables and the contribution of each to total variation is attempted also and thus the procedure that we are proposing can be viewed and evaluated against their results.

In Section II the method of simultaneous prediction is applied to the problem of selecting a subset of items from a given set. The criterion for a good selection is usually the degree of association of the items selected with a criterion variable exterior to the total set of items (See for instance the chapters by Sitgreaves in Solomon [1961]). However it may well happen that one does not have control over the criterion variable and we may know only that the total set is composed of items intended to measure different aspects of the same phenomenon. In this case it is natural to consider the whole observed set itself as the (multi-dimensional) criterion variable. Then the problem is transformed into one of predicting simultaneously all the items from a few of them, properly chosen. Given a subset of items we may apply one of two criteria for the validity of the given selection. The first one is that the subset must produce a first BSLP of the total set which explains a high percentage of the variance of the total set. The second criterion is that the subset must produce a complete sequence of such BSLP's.

At this point, however, an important difficulty arises. It is admitted that we can qualify for selection any given subset in terms of a certain criterion. But must we apply the criterion to all the possible subsets in order to choose the right one? Usually this implies an excessive amount of tedious work which borders on making attempts for solutions infeasible. In Solomon [1961], Elfving attacks a similar problem in the case of a known exterior criterion variable. He gives a geometrical method by which one avoids enumerating all the possible subsets to apply the criterion of selection. In Section II we suggest an enumeration of the subsets which makes it feasible to use the BSLP criterions or any other criterion $F(S)$, where F is any uni-dimensional function of the chosen subset S. This enumeration is a random one. Through the laws of simple random sampling we can compute how many selections need to be enumerated before one attains a probability α of having enumerated one of the selections which rank among the best. The number of selections desired is thus a function of a predetermined β, the proportion of all the selections we consider to be the best ones, or acceptable in some sense. (For instance those for which $F(S) \geq T$, where T is chosen so that the selections satisfying the inequality represent a proportion β of all the selections.) Though the enumeration is considerably reduced in most cases, the computations to be performed are still too numerous to be handled by a desk calculator. A digital computer is essential and thus a program that performs this task in the case where one applies the BSLP criterions (first BSLP, or all BSLP's) is referred to in this paper. For this program some typical output is given in Appendix B. In Section II results obtained from two sets of data with the help of this program are presented.

This procedure also permits a method of summarizing, or giving a score to a set of items. In particular, one might be interested in explaining the common variance (communalities) of a set of items by the observed values of the items which include errors due to specific variances, as in Section III. Also in that Section a method of estimating the common variance is given using only the already observed items. There, the criterion variables are the items after removal of specific variance. They are non-observable quantities, but we have the necessary information at our disposal.

In Part 3 of Section I the structure of simultaneous linear predictors is compared with the regression structure, and the resolution of the regression structure in terms of the simultaneous linear predictors is examined.

In Part 2 of Section I we examine a special case in which $\{Y_j\} \equiv \{X_i\}$, i.e., items and criterions are the same variables. This model is equivalent to Hotelling's development of "principal components." Here in addition to obtaining Hotelling's derivation from our main theorem, a new derivation which appears to be more direct and perhaps more elegant is given which does not make use of Lagrange multipliers.

We emphasize that the linear predictors treated here are the same kind of statistical objects as the "principal components" and the "canonical variates." They are underlying or latent variables and have an existence of their own in the realm of variables related to those the researcher is studying. To be of value as estimators, they need to be multiplied by a constant and, even then, they are not the best "simultaneous linear estimators." In other words, they do not, in general, minimize:

$$\sum_{j=1}^{m} E[Y_j - \mathcal{E}]^2$$

where $\mathcal{E}$ is a linear form in the X's. Here we are using the "squared loss" function. In Section I we derive those best estimators and note, in particular, that the mean of a test is a best estimator, in that sense, for the items of the test. Note then that the mean score on a test is good only in a very restricted and impractical situation. This should be intuitively so since it does not depend on intercorrelations.

The best estimators do not make use of the covariance matrix of the predictable parts of the items in the regression structure. They do not take into account, therefore, the relations between the items as the simultaneous linear predictors do. In very much the same way as we make use of the relations between the criterion variables in multiple regression, we make use of the relations between items in simultaneous prediction. Those ideas are clarified in Part 3 of Section I and in Section IV.

In Section IV we examine the problem of non-linear simultaneous prediction. In a special non-linear situation Sitgreaves solved the case where a uni-dimensional criterion is to be predicted or estimated for a given quadratic loss function. This is equivalent to the use of the index we have proposed. We give a general solution for the case where the criterion is vector-valued which puts her solution in a new light. To do this we again maximize

$$\sum_{j} \mathrm{Var}(Y_j)\rho^2_{Y_j S}$$

but without assuming S is a linear combination of the X_i's. However, we assume the vector X takes only a finite number of values. (Example: X_i dichotomous.)

Accompanying this paper, is a reference for and some output of a program for computations associated with the item selection procedure of Section II. Also in the appendix, we give proofs of the BSLP theorems.

SECTION I

1. Best Simultaneous Linear Predictors.

In the Introduction, the definition of a best simultaneous linear predictor is developed. This definition provides the index by which simultaneous linear predictors should be measured namely:

$$\sum_j \mathrm{Var}[Y_j]\rho^2_{Y_j S}$$

where the set of variables $\{Y_j\}$ are each to be predicted simultaneously by S, a linear function of the set of variables $\{X_i\}$. What is now required is a procedure for obtaining the coefficients of the variables in S. To accomplish this we now state the following main theorem:

Theorem 1.

"The best simultaneous linear predictor S_1 of a set of real random variables with zero means $\{Y_j\}$, $j=1,\ldots, m$ given a set of real random variables with zero means $\{X_i\}$, $i=1,\ldots, n$ is the linear combination of the X's:

$$S_1 = a_1' X$$

where a_1 is the vector associated with the largest characteristic root λ_1 of

$$R^{-1} T' T.$$

The matrix R is an $n \times n$ matrix whose elements are the covariances between X_i and X_k. It is not restrictive therefore to assume that

R is non-singular. The matrix T is also a covariance matrix between the Y_j's and the X_i's and is an $m \times n$ matrix.

The maximum value attained by the index $\sum_{j=1}^{m} \text{Var}\,[Y_j]\rho^2_{Y_j S_1}$ is λ_1. If the variance of S_1, equal to $a_1' R a_1$ is made unity by a proper normalization of a_1, the vector of regression coefficients of the Y's on S_1 is Ta_1.

We now state the complementary theorem:

Theorem 2.

"If from the original Y's ($\equiv Y^{(1)}$'s) one subtracts the random vector Ta_1S_1, the residual variables $\{Y_j^{(2)}\}$ can be considered for prediction, given the same X's. Denote the new T-matrix by T_2, i.e. $T \equiv T_1$. The best simultaneous linear predictor S_2 is in that case the linear combination

$$S_2 = a_2' X$$

Note: $\left[\begin{array}{ll} Y, X, S & \text{are random vectors} \\ S_i, X_i, Y_i & \text{are scalars.} \end{array}\right]$

where a_2 is the vector associated with the largest characteristic root λ_2 of $R^{-1} T_2' T_2$, λ_2 having the same property as before with regard to the maximum value attained by the index for the residual variables $Y_j^{(2)}$. Moreover, a_2 is the vector associated with the second largest root of $R^{-1} T' T$ which is also equal to λ_2. If a_2 is properly normalized as before, the vector of regression coefficients of the Y's and the $Y^{(2)}$'s, as well, on S_2 is Ta_2. Following the same method repeatedly, i.e., predicting successively

$Y - \sum_{j=1}^{i} Ta_j S_j$ for $i=1,\dots, p-1$, one obtains a sequence of simultaneous linear predictors $a_1' X, a_2' X,\dots, a_p' X$ with total variance explained: $\lambda_1 + \lambda_2 +\dots+ \lambda_p = \sum_{i=1}^{p} \lambda_i$, where

a_i are characteristic vectors of $R^{-1} T'T$

p = row-rank(T)[1]

λ_i are the ordered positive characteristic roots of $R^{-1} T'T$.

The final residual T-matrix, T_{p+1} is the zero matrix. Finally, the covariance or correlation matrix of the S_i's is the unit matrix, making the S_i's uncorrelated.

Proofs of Theorem 1 and Theorem 2 are given in appendix A of this report. Several corollaries can be deduced from the theorems and these are now presented.

Corollary 1.

If the set of Y_j's consists of Y_o alone, the classical multiple regression model, then T is a one-row matrix with row-rank $(T) = p = 1$. And $R^{-1} T' = a_1$, $TR^{-1} T' = \lambda_1$ since $R^{-1} T'T(R^{-1} T') = (TR^{-1} T')R^{-1} T'$. But from Theorem 1

$$\lambda_1 = \mathrm{Var}[Y]\rho^2_{YS_1}$$

---

[1] This is the same as rank (T) but emphasizes that T is not a square matrix.

so that the multiple correlation coefficient is

$$\rho_{YS_1} = \sqrt{\frac{\lambda_1}{\mathrm{Var}[Y]}} \cdot$$

Corollary 2.

Call $Ta_i = b_i$ the vector of regression coefficients of the Y_j's on S_i. By Theorem 2, the S_i's are uncorrelated. Suppose they are normalized with variance 1. Therefore, the variance explained by one S_i on some Y_j is b_{ij}^2 and the total variance explained by any number of S_i's on some Y_j is

$$\sum_{i \in A} b_{ij}^2$$

where A is the proper set for the indexes. This can be obtained from Corollary 1 where λ_1 becomes $\sum_{i \in A} b_{ij}^2$ by the fact that $R^{-1} = I$ and $T = (\mathrm{Cov}(Y_j, S_{i_1}), \ldots, \mathrm{Cov}(Y_j, S_{i_m})) = (b_{i_1 j}, \ldots, b_{i_m j})$ where $A = \{i_1, \ldots, i_m\}$.

Corollary 3.

If the Y_j's and the X_i's are identical, then

$$TR^{-1} T' = R^{-1} T'T = R$$

and the problem at hand is to find the characteristic vectors and roots of R. The resulting linear forms have been called "principal components" of a set of random variables by Hotelling [1934]. The "principal components" are, therefore, a particular case of BSLP's in which the index to be maximized reduces to $\sum_j \rho_{Y_j S}^2$ since all

$\mathrm{Var}[Y_j] = 1$ in that setting. However maximizing each $\sum_j \rho^2_{Y_j S_i}$ and realizing that its maximum is λ_i is exactly Hotelling's result.

2. Principal Components.

The last Corollary demonstrates that the "principal components" of a set of variables are a special case of simultaneous linear predictors. Though this is derived from Theorems 1 and 2, here is an alternative development for the essential part of the derivation which appears to be interesting.

Suppose that, given a set of random variables $\{X_i\}$, $i=1,\ldots, n$, one wants to maximize $\sum_{i=1}^{n} \mathrm{Var}[X_i]\rho^2_{X_i S}$, where S is of the form $S = a'X$. Then if $\underline{a}$ is a vector associated with the largest root λ_1 of R, we have a solution to the problem, and the quantity to be maximized becomes equal to λ_1.

Proof:

Let $\{\lambda_i\}$, $i=1,\ldots, n$, be the set of positive ordered roots of positive definite T. Let 0 be the matrix of associated, ordered, normalized, and orthogonal characteristic column vectors. Then $\sum_j \mathrm{Var}[X_i]\rho^2_{X_i S} = a'R^2a \,/\, a'Ra = \sum_i \lambda_i^2\, b_i^2 \,/\, \sum_i \lambda_i\, b_i^2$ where the vector

$b = 0'a$. Now $(\sum_i \lambda_i^2\, b_i^2 \,/\, \sum_i \lambda_i\, b_i^2) \leq \lambda_1 = \lambda_1^2\, b_1^2 \,/\, \lambda_1\, b_1^2$ by a simple

property of proportions, and equality is obtained when $b_1^2 = c$, a certain constant, and $b_i^2 = 0$, $i \neq 1$, or when $a = c$ (first column of 0).

Remark:

By the observation on p. 45 of Faddeeva [1959], with only minor changes in the proof, the results would hold when R is semi-definite and has roots of multiplicity greater than 1. The case $a'Ra = 0$ is assumed to be non-existent.

Corollary 4.

Suppose the second BSLP is defined as the linear form in the X_i's which explains the largest variance in the Y_j's and is also uncorrelated with the first BSLP. But then any predictor of the sort S'_2 would have the property that

$$\sum_j \mathrm{Var}(Y_j)\rho^2_{Y_j S'_2} = \sum_j \frac{\mathrm{Cov}^2(Y_j, S'_2)}{\mathrm{Var}(S'_2)}$$

$$= \sum_j \frac{\mathrm{Cov}^2(Y_j^{(2)}, S'_2)}{\mathrm{Var}(S'_2)} \leq \sum_j \frac{\mathrm{Cov}^2(Y_j^{(2)}, S_2)}{\mathrm{Var}(S_2)} = \sum_j \frac{\mathrm{Cov}^2(Y_j, S_2)}{\mathrm{Var}(S_2)}$$

since the middle quantity is maximized by S_2, the second BSLP, which moreover, is uncorrelated with the first BSLP. Therefore, the two definitions are equivalent.

Remark:

In the proofs of Theorems 1 and 2, to be found in the Appendix, we do *not* use any restrictive condition on the predictors (like the classical ones on the variance or on the vectors a_i), and we, therefore, do *not* use Lagrange multipliers at least explicitly.

3. BSLP's and the Regression Structure.

At this point the following question may arise naturally. How well do the BSLP's predict the Y_j's compared to a combination of all individual linear regressions of the Y_j's on the X_i's? The answer is that the BSLP's are in fact a resolution of the regression structure.

We have so far constructed an optimal linear structure of the Y's in terms of the S's,

$$Y = \sum_{j=1}^{p} Ta_j S_j + \epsilon = TAS + \epsilon \qquad [S_j \text{ scalar, } a_j \text{ vector}]$$

where ϵ is a random vector of residuals we cannot explain. This structure can be made explicit in terms of the X's, and we prove it is the regression structure,

$$Y = TR^{-1} X + \epsilon$$

where ϵ is the same as before. (The covariance matrix of the vector $TR^{-1} X$, $TR^{-1} T'$ plays a fundamental role in proving Theorems 1 and 2.) Indeed,

$$TAS = T(R^{-1} T' A^*)(A^{*\prime} TR^{-1})X$$

where A^* is the matrix of normalized characteristic vectors of $TR^{-1} T'$. This is true from the fact that

$$A = R^{-1}T'A^* \qquad \text{[Proof of Theorems 1, 2]}$$

and

$$S = A'X, \qquad \text{[Theorems 1, 2].}$$

It follows that

$$\begin{aligned} TAS &= A^* \operatorname{diag}(\lambda_1, \ldots, \lambda_n)\, A^{*\prime}\, TR^{-1} X \\ &= A^* \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})\, \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})\, A^{*\prime} TR^{-1} X \\ &= TR^{-1} X \end{aligned}$$

since $A^* \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})$ is an orthogonal matrix, recalling the characteristic vectors of $TR^{-1} T'$ are orthogonal, and that they have been normalized to have norm $1 / \lambda_i$, λ_i being the corresponding root.

Next we prove a theorem relating BSLP's to Principal Components from which some interesting corollaries are derived.

Theorem 3.

The BSLP's of a set of variables $\{Y_j\}$, given a set of variables $\{X_i\}$, are the Principal Components of the set of variables $\{Y^*_j\}$ where the vector

$$Y^* = TR^{-1} X$$

is the vector of linear regressions of the Y_j's on the X_i's.

Proof:

The vector of coefficients involved in each BSLP, S_i, is $R^{-1} T' a^*_i$. This comes from the proof of Theorems 1, 2. Then

$$S_i = (R^{-1} T' a^*_i)' X = a^{*\prime}_i TR^{-1} X .$$

If we recall that a^*_i is a characteristic vector of $TR^{-1} T'$ and note that $TR^{-1} T'$ is the matrix of covariances of the vector Y^*, the theorem follows.

Corollary 1.

In the case where $TR^{-1}X \equiv Y$ ("perfect regression") we can say the principal components of Y are the best simultaneous predictors of Y in terms of the criterion vector X. This can be viewed as a rationale for the classical use of principal components where we assume the criterions we want to consider explain (approximately) all the variance of the items. Moreover, if the simultaneous predictors involve only one criterion at a time, then the principal components of Y are identical to the criterions!

Corollary 2.

The first principal component of Y is the variable that maximizes $\sum_{j=1}^{m} \mathrm{Var}[Y_j]\rho^2_{Y_j Z}$ where Z is any real random variable. Indeed, one can consider a Z_o which maximizes the quantity over all possible variables. Then add $\{Y_j - C_j Z_o\}$ to Z_o, where C_j is the regression coefficient of Y_j on Z_o, to constitute a set of $(m+1)$ criterion variables. Through an orthogonal linear transformation reduce the $(Y_j - C_j Z_o)$'s to a set of uncorrelated variables. This latter set is still uncorrelated with Z_o. (Each $Y_j - C_j Z_o$ is uncorrelated with Z_o.) Then Z_o, together with the same set, constitutes a set of uncorrelated variables explaining the whole variance of Y, with positive covariance matrix R. And the first simultaneous linear predictor of Y in terms of criterions is obviously Z_o. Then apply the theorem and Corollary 1, to prove that Z_o is the first PC of the Y_j's.

4. Maximum Likelihood Estimation of the BSLP's Assuming Normality

If we assume the random vector $\begin{pmatrix} X \\ Y \end{pmatrix}$ is multinormally distributed with covariance matrix $U = \begin{pmatrix} R & T' \\ T & S \end{pmatrix}$ where R, T are as before and

$$S = [S_{ij}] = [\mathrm{Cov}(Y_i, Y_j)] ,$$

the matrix U gives the set of parameters of the distribution of the sample

$$\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix} \cdots \begin{pmatrix} X_p \\ Y_p \end{pmatrix}$$

of p observations, since we can assume without losing generality that the means are all zero. So far, we have really been concerned with the variables $\{X_i - \mu_i\}$ and $\{Y_j - \eta_j\}$, i.e., variables reduced by their means. Therefore in the estimation procedure we need to approximate the vectors μ and η, and we re-introduce them here in the population distribution to complete the set of parameters.

Reparametrize that distribution using the following one to one transformation $\mathcal{T}$: $(U, \mu, \nu) \xleftrightarrow{\mathcal{T}} (U^*, \mu, \nu)$ where U^* in the new parameter space consists of:

a.) The distinct characteristic roots and properly normalized characteristic vectors (so as to make them unique) of $W = R^{-1} T'T$ (assume the roots distinct).

b.) The distinct characteristic roots and properly normalized characteristic vectors (so as to make them unique) of $W' = T' TR^{-1}$.

(Given a) and b) one can reconstruct W using §32.2 in Faddeeva [1959]).

c.) T, S, μ, η.

d.) Some entries of R determined in the following way: Consider the matrix equation:

$$T'T = RW.$$

If one knows W and T, one has for each i, $i=1,\ldots, n$, a system of n equations for the n entries in the i^{th} row of R:

$$\sum_k r_{ik}w_{kj} = v_{ij}$$

where v_{ij} is the (i,j)-entry of $T'T$. But the rank of W may be smaller than n. We require in that case a number equal to $[n\text{-rank}(W)]$ of the elements in the vector r_i, say r_{ij_ℓ}, $\ell=1,\ldots, [n\text{-rank}(W)]$. The entries r_{ij_ℓ} constitute the last quantities needed to re-construct U from U^*. We have thus established the one to one transformation.

We therefore perform the same operations on the covariance matrix of the sample variables as we would on the covariance matrix of the population variables, to find the estimates of the roots and vectors of W. That follows from the transformation we have established which, by Lemma 3.2.3 in Anderson [1960] is the condition required for the maximum likelihood (ML) estimates of the quantities in (U^*, μ, ν) to be the image under $\mathcal{T}$ of the ML estimates of the quantities in (U, μ, ν). The latter estimates consist, by Theorem 3.2.1 in Anderson [1960] of the covariance matrix and the means of the sample variables.

To obtain ML estimates of the b_j's we proceed as before, using the sample equivalents of those quantities. These could be established similarly since the b_j's are uniquely determined by the BSLP and the matrix T.

5. Example of a Set of BSLP's for $m = 2$, $n = 2$.

Assume first $\mathrm{Var}(X_i) = \mathrm{Var}(Y_i) = 1$. We have:

$$R = \begin{pmatrix} 1 & \rho_{X_1X_2} \\ \rho_{X_1X_2} & 1 \end{pmatrix}, \quad T = \begin{pmatrix} \rho_{Y_1X_1} & \rho_{Y_1X_2} \\ \rho_{Y_2X_1} & \rho_{Y_2X_2} \end{pmatrix}$$

Then:

$$R^{-1}T'T = \frac{1}{1-\rho^2_{X_1X_2}} \begin{pmatrix} a & c \\ d & b \end{pmatrix}.$$

where:

$$a = \rho^2_{Y_1X_1} + \rho^2_{Y_2X_1} - \rho_{X_1X_2}(\rho_{Y_1X_1}\,\rho_{Y_1X_2} + \rho_{Y_2X_1}\,\rho_{Y_2X_2})$$

$$b = \rho^2_{Y_1X_2} + \rho^2_{Y_2X_2} - \rho_{X_1X_2}(\rho_{Y_1X_1}\,\rho_{Y_1X_2} + \rho_{Y_2X_1}\,\rho_{Y_2X_2})$$

$$c = \rho_{Y_1X_1}\,\rho_{Y_1X_2} + \rho_{Y_2X_1}\,\rho_{Y_2X_2} - \rho_{X_1X_2}(\rho^2_{Y_1X_2} + \rho^2_{Y_2X_2})$$

$$d = \rho_{Y_1X_1}\,\rho_{Y_1Y_2} + \rho_{Y_2X_2}\,\rho_{Y_2X_2} - \rho_{X_1X_2}(\rho^2_{Y_1X_1} + \rho^2_{Y_2X_1}).$$

The characteristic polynomial of $R^{-1}T'T$ is the determinant

$$\begin{vmatrix} \dfrac{a}{1-\rho^2_{X_1X_2}} - \lambda & \dfrac{c}{1-\rho^2_{X_1X_2}} \\ \dfrac{d}{1-\rho^2_{X_1X_2}} & \dfrac{b}{1-\rho^2_{X_1X_2}} - \lambda \end{vmatrix}$$

or:

$$\lambda^2 - \frac{a + b}{1-\rho^2_{X_1X_2}} \lambda + \frac{ab-cd}{1-\rho^2_{X_1X_2}} .$$

From which the roots of $R^{-1}T'T$ are:

$$\lambda_1 = \frac{a+b+\sqrt{(a-b)^2+4cd}}{2(1-\rho^2_{X_1X_2})}$$

$$\lambda_2 = \frac{a+b-\sqrt{(a-b)^2+4cd}}{2(1-\rho^2_{X_1X_2})} .$$

From the following system of equations:

$$\left(\frac{a}{1-\rho^2_{X_1X_2}} - \lambda\right) Z_1 + \frac{c}{1-\rho^2_{X_1X_2}} Z_2 = 0$$

$$\frac{d}{1-\rho^2_{X_1X_2}} Z_1 + \left(\frac{b}{1-\rho^2_{X_1X_2}} - \lambda\right) Z_2 = 0$$

we derive the characteristic vectors a_1 and a_2, these are the coefficients of the X's in the linear forms S_1 and S_2; $S_i = a_{i1}X_1 + a_{i2}X_2$:

$$a_{i1} = k_i , \; a_{i2} = \left(\frac{\lambda_i(1-\rho^2_{X_1X_2})-a}{c}\right) k_i$$

where the k_i are such that the S_i have variance equal to 1. Finally to obtain the vector of regression coefficients of the Y_i on each S_i, we multiply each a_i by the matrix T. Let us denote the results by b_i. These b_i are the "loadings" of the Y_i in each BSLP or S_i.

For a numerical computation we choose our data from Hotelling [1933]. These data were employed by Hotelling to illustrate the principal components solution he developed in that paper. Let

$$X_1 = \text{arithmetic speed}, \quad X_2 = \text{arithmetic power}$$

$$Y_1 = \text{reading speed}, \quad Y_2 = \text{reading power}$$

and

$$R = \begin{pmatrix} 1 & .594 \\ .594 & 1 \end{pmatrix}, \quad T = \begin{pmatrix} 2.64 & .081 \\ -.061 & .092 \end{pmatrix}$$

The BSLP's based on the X_i explain approximately 5% of the total variance of the Y_i. Naturally the total variance here is equal to 2. More precisely the variance explained by each S_i is given by the characteristic roots.

$$\lambda_1 = .0938 \quad \text{and} \quad \lambda_2 = .0142 .$$

Also we can write

$$S_1 = 1.24\, X_1 - .628\, X_2$$

$$S_2 = -.135\, X_1 + 1.08\, X_2$$

and

$$Y_1 = .276\, S_1 + .0518\, S_2$$

$$Y_2 = -.133\, S_1 + .108\, S_2 .$$

So it appears that speed minus power is the first BSLP and the opposite situation occurs for the 2nd one. On the other hand, it seems that arithmetic speed as opposed to arithmetic power is the best predictor of over all reading ability, $(\lambda_1 > \lambda_2)$, though it has a reverse relationship with reading power. Finally, it seems that reading speed is uncorrelated with the second BSLP and has only slight correlation with the first one.

We check our computations by a number of properties:

a) We have

$$.0939 = (b_{11})^2 + (b_{12})^2 = \lambda_1 \cong .0938$$

$$.0143 = (b_{21})^2 + (b_{22})^2 = \lambda_2 \cong .0142$$

b) The S_i are uncorrelated or

$$0 \cong a_{11}a_{21} + a_{12}a_{22} + \rho_{X_1X_2}(a_{11}a_{22} + a_{12}a_{21})$$

$$= .000216 .$$

c) The vectors b_i are orthogonal or

$$0 \cong (b_{11}b_{21} + b_{12}b_{22}) = -.0000672 .$$

6. Hints for Computations.

1. In order to compute the regression coefficients of the Y_j's on the S_i's one need not compute the roots and characteristic vectors of $R^{-1}T'T$ and then multiply them by the matrix T. It is sufficient to compute the roots and associated vectors of $TR^{-1}T'$ which give directly these coefficients since

$$b_i = \lambda_i a_i^*$$

where a_i^* is normalized so that

$$\sum_j a_{ij}^{*2} = \frac{1}{\lambda_i} .$$

2. In order to compute the proper b_i's from the characteristic vectors of $R^{-1} T' T$, it is not necessary to normalize the a_i's such that the variance of each S_i is 1. It is sufficient to compute Ta_i for even a non-normalized a_i, and, then, adjust b_i so that

$$\sum_j b_{ij}^2 = \lambda_i .$$

3. One has the choice of computing the roots and associated vectors of either $R^{-1} T' T$ or $TR^{-1} T'$. Though $TR^{-1} T'$ is more convenient since it is a symmetric matrix, it is possible that $R^{-1} T' T$ is of much smaller size as in the case of the example in the item selection problem treated in this paper. The dimension of $TR^{-1} T'$ is $m \times m$ and that of $R^{-1} T' T$ is $n \times n$, where m and n are as in Theorem 1. If there is a doubt as to the fact that the positive roots of $R^{-1} T' T$ are distinct, it is necessary to use $TR^{-1} T'$ for many methods of computing rely on the fact that the matrix used has linear elementary divisors. Symmetric matrices have that property (see Faddeeva [1959]). On the other hand, $R^{-1} T' T$ may not have that property if certain roots are multiple ones.

7. Simultaneous Linear Estimation.

In connection with what has been said in the introduction, let us prove the following theorem:

Theorem 3.1.

The best simultaneous linear estimator E of a set of real random variables $\{Y_j\}$ $j=1,\ldots, m$, based on a set of real random variables $\{X_i\}$, $i=1,\ldots, n$, is

$$E = b'X$$

where an element b_i of the vector b is the mean of the i^{th} column of the matrix of coefficients of the regression structure of the Y's on the X's. (Assume all the means are zero.)

Note: It is clear now that this estimator does not make use of the matrix $TR^{-1}T'$, the covariance matrix of the regression vector $TR^{-1}X$. For the best estimators all items to be predicted or estimated carry equal importance, which is contrary to the intuition that some items ought to be more important.

Proof: The best estimator, as defined before, is the linear form in the X_i's, $b'X$, which minimizes

$$\sum_j E(Y_j - b'X)^2$$

a quadratic form in the b_i's. Let us differentiate w. r. t. each b_ℓ, and equate the result to zero obtaining

$$\frac{\partial}{\partial b_\ell} \sum_j E(Y_j - \sum_i b_i X_i)^2 = 0$$

$$\sum_j E[\frac{\partial}{\partial b_\ell} (Y_j - \sum_i b_i X_i)^2] = 0$$

$$\sum_j E[(Y_j - \sum_i b_i X_i) X_\ell] = 0$$

$$\sum_j \left(E\, Y_j\, X_\ell - \sum_i b_i\, E\, X_i\, X_\ell\right) = 0$$

$$\sum_j \left(t_{j\ell} - \sum_i b_i\, r_{i\ell}\right) = 0$$

$$\sum_j t_{j\ell} - m \sum_i b_i\, t_{i\ell} = 0$$

$$\sum_i b_i\, r_{i\ell} = \frac{1}{m} \sum_j t_{j\ell}\, .$$

Or we may write

$$R\, b = \frac{1}{m} \begin{pmatrix} \sum_j & t_{j1} \\ & \cdot \\ & \cdot \\ & \cdot \\ \sum_j & t_{jn} \end{pmatrix},$$

$$b = \frac{1}{m} \sum_j R^{-1} \begin{pmatrix} t_{j1} \\ \vdots \\ t_{jn} \end{pmatrix}.$$

In other words, b is the mean vector of the vectors of regression coefficients corresponding to each Y_j, namely

$$R^{-1} \begin{pmatrix} t_{j1} \\ \vdots \\ t_{jn} \end{pmatrix}.$$

8. Regression on Orthogonalized Variables with the Use of Principal Components.

It is common in linear regression analysis to "orthogonalize" variables. This allows an economical approach to the computation of coefficients of regression and to the analysis when one wants as few variables as possible to be considered. A classical example is that of orthogonal polynomials. There are also other reasons, like the convenient theory for distributions and tests related to orthogonal variables.

Principal components (PC) are orthogonal variables in that sense and can be used in the same way. But they have a natural significance, however, that variables produced by other means may not have. Not only are they transformations of the regression variables into uncorrelated ones, but they are a most economical "image" of them if we consider the first PC, then the second one, etc. Therefore, assuming we want a regression of Y_o on the X_i's, we can tell whether the X_i's are chosen economically to explain Y_o, by investigating the relation between the first PC and Y_o, then between the second one and Y_o, etc. For example, we may discover that the first PC has little to do with Y_o, while the last one is almost identical to Y_o, meaning that Y_o and the X_i's as a whole, are in somewhat different universes even if the X_i's account for almost the total variance of Y_o. On the other hand, we discover which PC's are most related with Y_o, the compositions of which indicate a general direction for the search of new predicting variables. This is particularly true if some X_i's are highly related and are also related to Y_o. Their regression coefficients may be small because they "share" the same effect. (See for instance Anderson and Bancroft [1952] p. 202.) A PC is likely to gather them (specially if they have low correlation with other variables) and thus restore them as a high regression coefficient variable.

The method could be applied to studies as in French [1957] where an over-all rating of many teachers, Y_o, by their students is accompanied, in a questionnaire, by detailed appraisal items: $\{X_i\}$.

Let us compute the regression coefficients of Y_o, namely C_i, on S_i, the normalized PC's. We know

$$C_i = \mathrm{Cov}(Y_o, S_i)$$

from the fact that the S_i's are uncorrelated. But

$$\mathrm{Cov}(Y_o, S_i) = \sum_j a^*_{ij} \, \mathrm{Cov}(Y_o, X_j)$$

where a^*_i is the vector of the coefficients of the linear form involved in S_i, properly normalized. The variance explained by each S_i, is, of course, C_i^2, and the total variance explained by any number of S_i's is $\sum_{i \in A} C_i^2$ where A is the proper set of indexes.

SECTION II

1. Item Selection When No Latent Structure Is Assumed Known.

In Chapters 5, 6, and 7, in Solomon [1961], Elbving solves the problem of finding an optimal subset of items from a set of items to predict a uni-dimensional criterion Y. The quantity to be maximized is defined in terms of the common k factor structure of the items and the criterion. That is, each item X_i and the criterion Y are assumed to be linear functions of k common factors and naturally these factors are latent (non-observable) variables. The maximization procedure is also given and requires quite a number of mathematical considerations. We would like here to give an alternative development for both quantity to be maximized and the representation of the criterion, and also an alternative to the maximization procedure. Our criterion will be a vector one and will consist of the total set of items. (This appears to be a reasonable way of replacing a non-observable criterion.)

Suppose one has selected a set of n items out of N items and would like to know how well they predict the total set, and how they should be summarized so as to represent each item as well as possible. Obviously we can reduce those two problems to one by examining the Best Simultaneous Linear Predictor of the N items based on the n items. Then λ_1 is a measure of association and a criterion for the quality of the selection. ($\sum_{i=1}^{p} \lambda_i$ might be a better measure of association, divided by total variance or not; but, if one is planning to use a uni-dimensional score or index, the first root is the one which matters.)

This method could be applied, for instance, to studies as in Kahl and Davis [1956] where one might be interested in establishing a socio-economic status index based on, say, n out of 19 items. Of course, there may be one or many best selections in the finite set of $\binom{19}{n}$ possible selections, in that sense. With powerful means of computation one can examine all possibilities. With less powerful means one may compute how many random selections need to be examined, in order to choose the optimal one(s) with given high probability. This might decrease enormously the time of computation and make it practical on medium-sized computers.

Instead of computing the quantities associated with the first BSLP, one might in interested in the total set of BSLP's. In that case one need not extract all the roots, but merely consider the trace of $TR^{-1}T'$, or $R^{-1}T'T$, if one has to examine many selections. Other computations may be performed afterwards.

It is evident that one could also apply a criterion of non-linear prediction as developed in Section IV. But if one has to perform many selections, the computations involved may be much more tedious, and N may well be prohibitively large. In the next section, for a method of random selection, the number of selections necessary to attain a given degree of accuracy with at least a given probability with respect to a certain criterion is computed.

2. Random Selection of an Optimal Subset of n Items Out of a Set of N Items.

We have a criterion function F which takes on a specific real value for each of the $\binom{N}{n}$ subsets [distinct] containing exactly n

items. Call those subsets S_i, $i=1,\ldots,\binom{N}{n}$. Then the problem is to maximize the function $F(S_i)$ over all i; or we look for S^*, one of the S_i's, such that

$$\max_i F(S_i) = F(S^*) .$$

More generally let us assume the values taken by F (some possibly equal) have been ordered from the largest to the smallest: $F_1, F_2,\ldots,F_{\binom{N}{n}}$. There is a corresponding ordering for the subsets S_i: $S_{i_1}, S_{i_2},\ldots,S_{i_{\binom{N}{n}}}$, i.e.,

$$F(S_{i_1}) = F_1 \geq F(S_{i_2}) = F_2 \geq \cdots \geq F(S_{i_{\binom{N}{n}}}) = F_{\binom{N}{n}} .$$

We note that $S_{i_1} = S^*$. Suppose now that any one of the k best subsets is suitable for our purposes (those subsets which give the k largest values for F). This means that we want to find any subset which lies in the class of subsets representing the top proportion, say $\beta = \frac{k}{\binom{N}{n}}$, of the total number of them.

It is clear that if we choose each subset with equal probability, the probability of picking up one in which we are interested (one in the top β-proportion) is equal to β. Sampling repeatedly with replacement, the probability α of not choosing one of the optimal sets in m trials is,

$$(1-\beta)^m = \alpha . \tag{1}$$

In other words α is the probability of no success in m independent Bernoulli trials with parameter β.

For a given β, we can determine the value of m which will insure a pre-assigned value of α. Thus we are saying that if one is willing to accept a selection which ranks in the top 5%, for example, with a probability α of being in error as to the choice, one can compute a suitable m.

Let us now transform equation (1). From (1) we have

$$\log[(1-\beta)^m] = \log \alpha$$

or

$$m = \frac{\log \alpha}{\log(1-\beta)} \quad (2)$$

For given α, β, the value of m in equation (1) generally will not be an integer. So, in practice, m will be the smallest integer greater than or equal to $\frac{\log \alpha}{\log(1-\beta)}$. Let us denote that integer by

$$\left[\frac{\log \alpha}{\log(1-\beta)}\right] .$$

Next we give a table of values for $\left[\frac{\log \alpha}{\log(1-\beta)}\right] = m$ using 5-place logarithms.

1. Table of $m = \left[\frac{\log \alpha}{\log(1-\beta)}\right]$, for values of α and β.

β \ α	.20	.10	.05	.01	.001	.0001
.20	8	11	14	21	31	42
.10	16	22	29	44	66	88
.05	32	45	59	90	135	180
.01	161	230	299	459	689	918
.001	1,626	2,326	3,026	4,652	6,977	9,303
.0001	17,475	25,000	32,526	50,000	75,000	100,000

Note 1: In a programmed computer one need not keep all the values $F(S_i)$ given by the successively chosen S's. One need retain only the largest value obtained through the last inspection, as well as an indexing of the items of the subset corresponding to that value.

Note 2: The more equivalent the items are suspected to be with respect to the criterion function, the larger β should be chosen, and the smaller α, for a desired m. If things are to be done economically, in general, one should insist more on reliability (small α) and less on accuracy (large β), since the function $\frac{\log \alpha}{\log(1-\beta)}$ is much more sensitive to β than to α.

3. Worked Out Examples of Item Selection.

In Kahl and Davis [1956] and French [1957] item selection is attempted in the former for socio-economic data and in the latter for an analysis of teacher ratings. Let us use the random selection technique described above, and apply the BSLP computation to each selection: as a first step we get the BSLP of the total set of items given the selection, and its contribution to the variance of the total set of items from all BSLP's based on the selection. In other words for each selection we compute the largest eigenvalue of R^{-1} T'T (contribution to variance of first BSLP), an associated eigenvector (coefficients of the items selected in the linear form $S_1 = a_{11}X_{i_1} + \dots + a_{1n}X_{i_n}$ where $X_{i_1}, \dots, X_{i_n}$ are the items selected), and the trace of R^{-1} T'T (total contribution of all items in selected subset or all BSLP's for this subset to the total variance of the total set of items.) We have seen that the last quantity is also the cumulative variance explained by all linear regressions of all the items on the subset selected. Next we give tables of the maximum value attained by λ_1, the largest root of R^{-1} T'T, and the maximum value attained by trace (R^{-1} T'T) when R^{-1} T'T varies over 100 selections of items yielding 100 subsets of items and the associated R, T. Moreover, we have 100 selections for each size of the selected subset from 2 to 10 items (a total of nine hundred selections of varying size). We also give the ratio of the maximum trace over the dimension of the source matrix (matrix of correlations of all items). The dimension of the source matrix is the total variance of all items since they are normalized to have variance 1.

In Kahl and Davis [1956] the dimension of U the source matrix is 19×19. In French [1957] it is 41×41. Finally we give the ratio of the maximum λ_1 over the largest root of U which corresponds to the variance explained by the first BSLP based on the total set of items, in other words the first PC. For the former example, this is 9.63 approximately (computed using the "centroid method"). In the latter example it is 22.2465.

TABLE 1 : U : 19 × 19 . */

No. of Items	(1) Max $Tr(R^{-1} T'T)$	(2) Max $\lambda_1(R^{-1} T'T)$	(3) $(1)/19$	(4) $(2)/9.63$
2	10.25863	8.96855	.54	.93
3	11.84774	9.40517	.62	.98
4	13.34550	9.55025	.70	.99
5	14.40331	9.68343	.76	1.00
6	15.62095	10.03782	.82	1.04
7	18.79842	9.84465	.99	1.02
8	18.30988	10.19383	.96	1.06
9	18.32662	10.60689	.96	1.10
10	18.91390	10.20557	.99	1.06

---

*/ The correlations used are tetrachoric correlations leading here to approximate results. Entries in col (4) over one are non-acceptable and are probably due to the fact that U is possibly not positive semi-definite. The use of tetrachoric correlations provides a matrix with properties not generally associated with covariance matrices.

TABLE 2 : $U : 41 \times 41$.

No. of Items	(1) Max $Tr(R^{-1}\ T'T)$	(2) Max $\lambda_1(R^{-1}\ T'T)$	(3) $^{(1)}/_{41}$	(4) $^{(2)}/_{22.2465}$
2	21.43907	19.74671	.52	.89
3	24.49352	20.75624	.60	.93
4	25.76339	20.99744	.63	.94
5	27.27984	21.31813	.67	.96
6	28.11818	21.38823	.69	.96
7	29.26203	21.55163	.71	.97
8	29.78843	21.77101	.73	.98
9	30.96915	21.82832	.76	.98
10	31.76981	21.93530	.77	.99

While the trace in Table 1 attains a quasi absolute maximum at 7 items the trace in Table 2 ceases to increase very sharply after 9 items and the rate of increase changes to a smaller one after 5 items. For the root (λ_1) it seems that in both cases the rate of increase changes after 3 items, if we neglect the apparent erratic behavior of the root in Table 1, due to the type of correlations (tetrachoric) used.

One may use the "trace" criterion or the "root" criterion, as already explained. If one needs to summarize the items in a real-valued predictor it is desirable to use the root as a criterion. We then obtain the following selection of 3 items (see listing in Appendix B):

In Kahl and Davis [1956]: Items 11, 16, 2 ;

In French [1957]: Items 5, 25, 31 .

4. Items Selected as Oblique Components.

It is interesting to parallel column (1) in Table 2, to the cumulative contribution to total variance given in French [1957] for the first 8 Principal Components.

	1	2	3	4
Components	22.2465	22.8868	28.0456	29.7032
Items		21.43907	24.49352	25.76339
	5	6	7	8
Components	31.0649	32.1476	33.1044	33.8832
Items	27.2798	28.11818	29.2620	29.78843

We draw the conclusion that we have here a method of finding oblique (correlated) components, or factors, which are the most easily identifiable of all, since they are some of the items!

Note: The computations of this section have been performed with the help of a program which is referenced in Appendix B.

SECTION III

1. Summarizing Correlated Items With Specific Variances.

Consider a set of items of the form

$$X_i = F_i + \epsilon_i$$

where the ϵ_i's are uncorrelated with the F_i's and with each other. The PC's of such a set of items are not, in general, the BSLP's of the F_i's based on the X_i's. Let us examine how much variance each S_j explains in the F_i's. The variance explained in each F_i is the square of $\mathrm{Cov}(F_i,S_j)$. This can be seen in the following way. The variance explained is $\mathrm{Var}(F_i)\rho^2_{F_iS_j}$. But

$$\mathrm{Var}(F_i)\rho^2_{F_iS_j} = \frac{\mathrm{Var}(F_i)\mathrm{Cov}^2(F_i,S_j)}{\mathrm{Var}\ F_i \mathrm{Var}\ S_j} = \mathrm{Cov}^2(F_i,S_j) \text{ if } \mathrm{Var}\ S_j = 1.$$

Now

$$\mathrm{Cov}(F_i,\ S_j) = \mathrm{Cov}(F_i + \epsilon_i,\ S_j) - \mathrm{Cov}(\epsilon_i,\ S_j) .$$

But

$$\mathrm{Cov}(\epsilon_i,\ S_j) = a_{ji}\ \sigma_i^2$$

where σ_i^2 is $\mathrm{Var}(\epsilon_i)$, and a_{ji} is the coefficient of X_i in S_j. And

$$\mathrm{Cov}(F_i + \epsilon_i,\ S_j) = \lambda_j\ a_{ji}$$

since $X_i = \sum_j \lambda_j\ a_{ji}\ S_j = \sum_j b_{ji}\ S_j$ [PC's leave no residual variance unexplained.] Now

$$\begin{aligned} \mathrm{Cov}^2(F_i, S_j) &= (\lambda_j\, a_{ji} - a_{ji}\, \sigma_i^2)^2 \\ &= (a_{ji})^2(\lambda_j - \sigma_i^2)^2 \, . \end{aligned}$$

So the common variance explained in each X_i is the variance explained in each X_i, namely $\lambda_j^2\, a_{ji}^2$, reduced by the multiplicative quantity

$$\left(\frac{\lambda_j - \sigma_i^2}{\lambda_j}\right)^2$$

which is close to 1 if σ_i^2 is small with respect to λ_j. The total variance explained on all X_i's is $\sum_i (\lambda_j\, a_{ji})^2 = \lambda_j$ and the total common variance explained is

$$\sum_i \left[\lambda_j\, a_{ji}\, \frac{(\lambda_j - \sigma_i^2)}{\lambda_j}\right]^2 \leq \lambda_j \, .$$

In Section 8 of Hotelling [1933] there is an attempt to find a good approximation of the principal components of a set of true "test scores," based on the test scores which are a combination of true test scores and "chance errors".

Since the principal components are nothing but best simultaneous linear predictors based on the true test scores, instead of approximating them, one might rather find directly best simultaneous linear predictors (for true scores) based on test scores. In this case the T-matrix is R where the diagonal has been reduced according to specific variances.

The method can be applied to summarize tests where the items are inter-related and where the errors of measurement or other specific variables have known variances.

2. Approximation for the Values of the Common Variances of Items when Those Variances are Unknown and not Obtainable by Direct Observation

As before, let the items X_i be of the form

$$X_i = F_i + \epsilon_i$$

where the ϵ_i's are uncorrelated with the F_i's and with each other. In what has been said in the preceding parts of this section, the quantities $\mathrm{Var}(F_i)$ play a crucial role. But, unfortunately, the F_i's are not directly observable, and it might be undesirable to construct an experiment in order to get a new X_i':

$$X_i' = F_i + \epsilon_i' ,$$

where ϵ_i' is uncorrelated with ϵ_i and $\mathrm{Cov}(X_i', X_i) = \mathrm{Var}(F_i)$. For example one may repeat a test to measure the amount of error due to extraneous factors which arise in the administration of the test.

If this is the case, we suggest that the following procedure be applied. For each item X_j find the linear regression on all the other items. The variance of the linear combination of X_i's, $i \neq j$, are an approximation for $\mathrm{Var}(F_i)$ if the F_i's are of the form

$$F_i = \sum_{k=1}^{p} a_k G_k$$

or, in other words, where the ϵ_i's include any specific component of the X_i's. The rationale of the procedure is that the ϵ_i's will be absorbed in the residuals of the regressions (they do not appear in the linear form $\sum_{i \neq j} C_i X_i$), and that the weighted sum of the ϵ_i's, $i \neq j$, included in the regression will have small variance, making F_i approximately equal to $\sum_{i \neq j} C_i X_i$. Though this is only a conjecture in the general case, we can prove the same in a particular case where

$$F_i = a_i F$$

or in the so-called "single factor" structure. It is understood that there are ways of calculating the a_i's, but we nevertheless apply our procedure rather artificially in order to give some justification to the conjecture for the general case.

We want to compute:

$$X_i' = t_i' R_i^{-1} X$$

namely, the linear regression of X_i on the rest of the X_j's, where $t_i' = (a_i a_1, \dots, a_i a_{i-1}, a_i a_{i+1}, \dots, a_i a_n)$

$$R_i = \begin{array}{llllll}
1, & a_1a_3, \dots, & a_1a_{i-1}, & a_1a_{i+1}, \dots, & a_1a_n & \\
a_2a_1, & 1, \dots, & a_2a_{i-1}, & a_2a_{i+1}, \dots, & a_2a_n & \\
\dots & \dots & \dots & \dots & \dots & \\
a_{i-1}a_1 & \dots & \dots & \dots & \dots & \dots \\
a_{i+1}a_1 & \dots & \dots & \dots & \dots & \dots \\
\dots & \dots & \dots & \dots & \dots & \dots \\
a_n a_1 & \dots & & \dots & \dots & 1
\end{array}$$

$$X' = (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$$

For simplicity we have assumed $Var(X_j) = 1$. Without loss of generality we put $X_i = X_1$, and then:

$$R_1^{-1} = \left[\begin{pmatrix} 1 - a_2^2 & 0 & . & . & . & 0 \\ 0 & 1 - a_3^2 & & & & . \\ 0 & 0 & 1 - a_4^2 & & & . \\ . & . & . & . & \ddots & . \\ & & & & . & 0 \\ 0 & . & . & . & . & 0 \quad 1 - a_n^2 \end{pmatrix} + aa'\right]$$

where $a' = (a_2, a_3, \ldots, a_n)$, or

$$R_1^{-1} = (D_1 + aa')^{-1}$$

$$= D_1^{-1} - D_1^{-1}\, a(1 + a'\, D_1^{-1}\, a)^{-1}\, a'\, D_1^{-1} .$$

Using a known equality, we get

$$R_1^{-1} = D_1^{-1} - D_1^{-1}\, a(1 + \sum_{i \neq 1} \frac{a_i^2}{1 - a_i^2})^{-1}\, a'\, D_1^{-1} .$$

If we let $d_1 = 1 + \sum_{i \neq 1} \frac{a_i^2}{1 - a_i^2}$, we obtain

$$R_1^{-1} = D_1^{-1} - D_1^{-1}\, a\, d_1^{-1}\, a'\, D_1^{-1}$$

$$R_1^{-1} = D_1^{-1} - d_1^{-1} \left\{ \frac{a_i\, a_j}{(1 - a_i^2)(1 - a_j^2)} \right\}, \; i, j \neq 1$$

where the quantity in brackets is a matrix. Then in matrix form:

$$R_1^{-1} = \{s_{ij}\}$$

where

$$s_{ij} = \begin{cases} - d_1^{-1} \dfrac{a_i\ a_j}{(1 - a_i^2)(1 - a_j^2)} & \text{for } i \neq j \\ & i,\ j \neq 1 \\ \dfrac{1}{1 - a_i^2} - d_1^{-1} \dfrac{a_i^2}{(1 - a_i^2)^2} & \text{for } i = j \\ & i,\ j \neq 1 . \end{cases}$$

Applying this result to

$$X_1' = t_i'\ R_i^{-1}\ X$$

we get:

$$x_1' = \sum_{j=2}^{n} d_1^{-1} \frac{a_1\ a_j}{1 - a_j^2}\ X_j$$

or:

$$X_1' = \sum_{j=2}^{n} d_1^{-1} \frac{a_1\ a_j}{1 - a_j^2}\ (a_j\ F + \sqrt{(1 - a_j^2)}\ \delta_j)$$

where δ_j has variance 1, and we also assume F has variance 1. Now $(1 - a_j^2)\ \delta_j = \epsilon_j$, thus:

$$X_1' = \left(\sum_{j=2}^{n} d_1^{-1} \frac{a_1\ a_j^2}{1 - a_j^2} \right) F + \sum_{j=2}^{n} d_1^{-1} \frac{a_1\ a_j}{\sqrt{1 - a_j^2}}\ \delta_j$$

and

$$\mathrm{Var}(X_1') = a_1^2(1 - d_1^{-1})$$

since the δ_j's and F are uncorrelated. Or, in general:

$$\mathrm{Var}(X_1') = a_i^2(1 - d_i^{-1})$$

where

$$d_i = 1 + \sum_{j \neq i} \frac{a_j^2}{1 - a_j^2} \ .$$

The quantity $(1 - d_i^{-1})$ is close to 1 if d_i is large. This is the case if the a_i's are large or if we have a large number of items. The larger the a_i's, the smaller the number of items required. For a given set of items, the small a_i's will be better approximated, which is a desirable property. We have given an example taken from Hotelling [1933] where the a_i^2's are as follows in the left column of the next table. (In reality, those a_i^2's are associated with a multiple factor structure.) On the right hand side we give the approximate values obtained through the computation of $a_i^2(1 - d_i^{-1})$.

True a_i^2 values	Approximate values
.9197	.8735
.8942	.8564
.9083	.8660
.5639	.5452

The approximation is shown to be good enough even for a very small number of items. It could be, of course, considerably improved by the addition of a small number of items (say 5 to 10).

We believe this result could be generalized to a multiple factor structure. This should be the subject of some further work.

Remark: Before using any X_i' instead of X_i, one must remember the new specific variables (e.g. errors associated with each X_i') are correlated. Therefore, a linear sum involving every X_i' may have an equal or greater error variance than the same linear sum involving every X_i. But an X_i' could be used individually as an estimate of the informative part of an item as opposed to its specific "noise" component. In the single factor case this estimate can be shown to be much better than X_i. Indeed

$$X_i' = a_i F - a_i d_i^{-1} F + d_i^{-1} \sum_{j \neq i} \frac{a_i a_j}{\sqrt{1 - a_j^2}} \delta_j .$$

The variance of the last two terms on the RHS of the equation which represent the error in estimating $a_i F$ is $\frac{a_i^2}{d_i}$. Using the same example as before, where $(1 - a_i^2)$ represents the variance of the error of estimating $a_i F$ by X_i, we give a table comparing $1 - a_i^2$ to a_i^2/d_i.

$1 - a_i^2$	a_i^2/d_i
.0803	.0462
.1158	.0378
.0917	.0423
.4361	.0187

One might be tempted to consider the first two terms of the RHS of the last equation as the informative part of X_i' and regard the last term as a specific component having a smaller variance than that of X_i. The variance of the component is $a_i^2(\frac{1}{d_i} - \frac{1}{d_i^2})$, a value very close to $\frac{a_i}{d_i}$. But once again, in order to use the information, one has to construct a predictor involving a linear sum of every X_i' for instance, and due to the fact that the new specific components are correlated, one does not expect a sure gain.

SECTION IV

1. Non-Linear Prediction Given a Set of Items: A Problem Proposed by Sitgreaves.

In Chapter 2 of Solomon [1961], Sitgreaves assumed a set of items $\{X_i\}$ possesses a certain structure called a Guttman scale. From those items she desires a good estimate $\hat{Y} = f(X_1,\ldots,X_n)$ of a uni-dimensional criterion variable Y. The estimate is not assumed to be linear, but is expected to satisfy the property that $\hat{Y}$ minimize:

$$E(\hat{Y}-Y)^2 .$$

Let us develop this estimate in a more general situation, that is without the restriction of a Guttman scale. Later we will permit Y to be vector-valued. We know:

$$E(\hat{Y}-Y)^2 = E\{E[\hat{Y}-Y)^2|X_1,\ldots, X_n]\} .$$

Therefore, if we minimize:

$$E[(\hat{Y}-Y)^2 \mid X_1,\ldots, X_n]$$

for each given value of the vector X we also minimize $E(\hat{Y}-Y)^2$ which is an average of those quantities. But we may write

$$E[(\hat{Y}-Y)^2|X_1,\ldots, X_n] = Var(Y|X_1,\ldots,X_n) + E^2(Y|X_1,\ldots,X_n) + \hat{Y}^2 - 2\hat{Y}\, E(Y|X_1,\ldots, X_n) .$$

The first term on the RHS is the conditional variance of Y, a term independent of $\hat{Y}$, and the sum of the other terms can be made zero by putting:

$$\hat{Y} = E(Y|X_1, \ldots, X_n) .$$

Thus the conditional expectation of Y is the desired predictor. And of course,

$$E(\hat{Y}-Y)^2 = E[\mathrm{Var}(Y|X_1, \ldots, X_n)]$$

which enters as the significant term in a squared correlation ratio.

Moreover, if $\hat{Y} = E(Y|X_1, \ldots, X_n)$ the regression coefficient of Y on $\hat{Y}$ is obviously 1. Therefore from linear regression theory, we have:

$$E(\hat{Y}-Y)^2 = E\ \mathrm{Var}(Y|X_1, \ldots, X_n) = \mathrm{Var}(Y) - \mathrm{Var}(Y)\rho^2_{Y\hat{Y}}$$

and the variance "explained" is $\mathrm{Var}(Y)\rho^2_{Y\hat{Y}}$, which relates the squared loss function to the idea of a correlation coefficient in an inescapable way. As a matter of fact, $\hat{Y} = E(Y|X_1, \ldots, X_n)$ is the real random variable that has maximum correlation coefficient with Y, among all those that are functions of the X_i's. We prove this by contraposition in the following manner. [1] Suppose there is a function of the X_i's, say $\hat{Z}$, that has a higher correlation with Y than $\hat{Y}$. Without changing the value of $\rho_{\hat{Z}Y}$ one can assume that $\mathrm{Var}(\hat{Z}) = \mathrm{Var}(\hat{Y})$ and also that Y, therefore $\hat{Y}$, and $\hat{Z}$ have means equal to zero. (If not, multiply Z by a constant.) Then, by simple algebraic manipulations, we obtain

---

[1] This is also a direct consequence of the discussion of simultaneous prediction in the Introduction.

$$E(Z\text{-}Y)^2 = \sqrt{\mathrm{Var}(\hat{Z})\ \mathrm{Var}(Y)}\ \left[\frac{\sqrt{\mathrm{Var}(Y)}}{\sqrt{\mathrm{Var}(\hat{Z})}} - 2\rho_{\hat{Z}Y} + \frac{\sqrt{\mathrm{Var}(\hat{Z})}}{\sqrt{\mathrm{Var}(Y)}} \right]$$

$$\leq \sqrt{\mathrm{Var}(Y)\ \mathrm{Var}(\hat{Y})}\ \left[\frac{\sqrt{\mathrm{Var}(Y)}}{\sqrt{\mathrm{Var}(\hat{Y})}} - 2\rho_{\hat{Y}Y} + \frac{\sqrt{\mathrm{Var}\ \hat{Y}}}{\sqrt{\mathrm{Var}(Y)}} \right] = E(\hat{Y}\text{-}Y)^2$$

which is a contradiction if $\hat{Z}$ is a function of the X_i's.

Applying these results to items with a Guttman scale we obtain the predictor and the corresponding index given by Sitgreaves. With a squared loss function the conditional expectation of Y given $X_1, \ldots, X_n$ is the best possible predictor among predictors of all forms, and is therefore better than any linear predictor.

But if one assumes a vector criterion Y instead of a uni-dimensional criterion, an approach for non-linear prediction for this situation could borrow from this development. We have just seen that in non-linear prediction for a uni-dimensional Y we maximize an equivalent of the quantity

$$\sum_j \mathrm{Var}\ Y_j\ \rho^2_{\hat{Y}_o Y_j} ,$$

using a given predictor $\hat{Y}_o$. When we discussed simultaneous prediction in the Introduction it was not assumed that $\hat{Y}_o$ was a linear predictor. Let us try here to solve the problem in the general non-linear case with a vector criterion, assuming however a more restrictive type of item, with discrete vector responses, for example, those obeying a Guttman scale, where the vector-valued item response X has one of the possible values

$$X_{(k)} = (1, 1, 1, \ldots, 1, 0, \ldots, 0) ,$$

that is the first k item variables are 1 and the rest are 0. Now assume

$$\Pr(X_{(k)}) = P_k .$$

Let the predictor $\hat{Y}_o$ take the corresponding values

$$\hat{Y}_o(X_{(k)}) = \hat{y}_k$$

with the same probability P_k. Assume Y has mean zero. We have

$$\sum_j \operatorname{Var}(Y_j)\, \rho^2_{\hat{Y}_o Y_j} = \frac{\sum_j \left(\sum_k (\hat{y}_k - \sum_r P_r \hat{y}_r)\, P_k \nu_{kj} \right)^2}{\sum_k (\hat{y}_k - \sum_r P_r \hat{y}_r)^2\, P_k} = \lambda(\hat{y})$$

where $\nu_{kj} = E(Y_j | X_{(k)})$ and $\hat{y}$ is the vector of the $\hat{y}_k$'s. Before trying to maximize $\lambda(\hat{y})$ let us consider

$$\lambda^*(S) = \frac{\sum_j (\sum_k S_k\, P_k\, \nu_{kj})^2}{\sum_k S_k^2\, P_k} .$$

It is clear that

$$\max_{\hat{y}} \lambda(\hat{y}) \leq \max_S \lambda^*(S)$$

since

$$\lambda(\hat{y}) = \lambda^*(S(\hat{y})) .$$

Let us find $\max_S \lambda^*(S)$. We have already proved (Lemma 2 in the Appendix) in a more general case that ratios of quadratic forms of that sort have an absolute maximum on R_m excluding the origin. (We assume not all P_k's are zero). Differentiate λ^* with respect to each S_ℓ using the relation for the derivative $(\frac{u}{v})'$ of a ratio:

$$(\frac{u}{v})' = 0 \; ; \quad u' = \frac{u}{v} v' \; .$$

We have

$$\sum_j \sum_k S_k \, P_k \, \nu_{kj} \, \nu_{\ell j} = \lambda^* \, S_\ell \, , \quad \ell = 1, \ldots, m$$

if we equate the partial derivatives to 0. In other words:

$$\sum_k{}' \, (\sum_j{}' \, \nu_{kj} \, \nu_{\ell j}) \, P_k \, S_k = \lambda \, S_\ell \, , \quad \text{for all} \quad \ell$$

or in matrix notation:

$$N \, S = \lambda^* \, S$$

where $n_{\ell k} = P_k \, \nu_k' \, \nu_\ell$, ν_k and ν_ℓ being the vectors of the ν_{kj}'s and the $\nu_{\ell j}$'s . A solution of the problem is then to choose S a characteristic vector associated with the largest root λ_1^* of N and

$$N \, S = \lambda_1^* \, S$$

the maximum attained for λ^* being λ_1^*, since one can verify that $\lambda^*(S) = \lambda_1^*$ for that choice of S. (Just express the ratio λ^* in matrix notation.) Now we prove that

$$\sum_k P_k \, S_k = 0 \; .$$

Indeed from the equation: $NS = \lambda_1^* S$ we get

$$\sum_{\ell} P_{\ell} \sum_{j} \sum_{k} S_k P_k \nu_{kj} \nu_{\ell j} = \lambda_1^* \sum_{\ell} P_{\ell} S_{\ell}$$

$$= \sum_{j} \sum_{k} S_k P_k \nu_{kj} \quad (\sum_{\ell} P_{\ell} \nu_{\ell j}) = 0$$

since by assumption $\sum_{\ell} P_{\ell} \nu_{\ell j} = E(Y_j) = 0$. Using this result we obtain

$$\lambda(S) = \lambda^*(S) = \lambda_1^*$$

if S is the characteristic vector associated with λ_1^*. Recalling what has been said before, it means that

$$\lambda_1^* = \max_{\hat{y}} \lambda(\hat{y}) = \lambda_1$$

and that this value is attained for $\hat{y}$ satisfying

$$N \hat{y} = \lambda_1 \hat{y}$$

where λ_1 is the largest root of N. It follows from the lemma we prove next, that the roots of N are non-negative and that the largest root is also the largest one in modulus. To apply the lemma one must note that

$$N = P^{-1} L$$

where $(\hat{y} L y)$ and $(\hat{y} P y)$ are the quadratic forms in

$$\lambda^*(\hat{y}) = \frac{\hat{y} L y}{\hat{y} P y}$$

and that P as well as P^{-1} is positive definite while L is positive semi-definite.

Lemma: Let $A: n\mathrm{x}n$ be a positive semi-definite matrix and $B: n\mathrm{x}n$ be a positive definite matrix. Then BA is a positive semi-definite matrix.

Proof: Consider the ratio of quadratic forms

$$\frac{c'Ac}{c'B^{-1}c} \geq 0$$

and let λ_i be a root of BA and c_i a corresponding vector; we have

$$BAc_i = \lambda_i c_i$$

or

$$Ac_i = \lambda_i B^{-1} c_i$$

and

$$\lambda_i = \frac{c_i' A c_i}{c_i' B^{-1} c_i} \geq 0 \ .$$

Now, as in Sitgreaves' development, assume there is an underlying real random variable, say Z, jointly normally distributed with the Y_j's (all variances equal to 1) and such that

$$X_i = 1 \text{ if } Z \geq \alpha_i \ , \ \alpha_i \geq \alpha_{i-1}, \text{ and } X_i = 0, \text{ otherwise.}$$

This structure produces a Guttman scale. Then, from Sitgreaves:

$$(1) \qquad \nu_{kj} = E(Y_j | X_{(k)}) = -\rho_{Y_j Z} \frac{\varphi(\alpha_{k+1}) - \varphi(\alpha_k)}{\Phi(\alpha_{k+1}) - \Phi(\alpha_k)}$$

where φ is the normal (0,1) density and Φ the corresponding distribution function and the $X_{(k)}$'s are ordered and defined as before. Then

$$n_{\ell k} = P_k \nu_k' \nu_\ell = \left(\sum_j \rho^2_{Y_j Z}\right) \left\{ \frac{[\varphi(\alpha_{k+1}) - \varphi(\alpha_k)][\varphi(\alpha_{\ell+1}) - \varphi(\alpha_\ell)]}{\Phi(\alpha_{\ell+1}) - \Phi(\alpha_\ell)} \right\}$$

and for the characteristic vector corresponding to the largest root of N choose the elements

$$y_k = \frac{\varphi(\alpha_{k+1}) - \varphi(\alpha_k)}{\Phi(\alpha_{k+1}) - \Phi(\alpha_k)} .$$

This vector depends only on the item distribution. This is natural, since the different Y_j's have a relation to the items only through Z. The largest root of N is:

$$\left(\sum_j \rho^2_{Y_j Z}\right) \left[\sum_k \left\{ \frac{[\varphi(\alpha_{k+1}) - \varphi(\alpha_k)]^2}{\Phi(\alpha_{k+1}) - \Phi(\alpha_k)} \right\} \right]$$

since N being obviously of dimension 1 has only one positive root, and that root corresponds to the characteristic vector chosen. And

$$b_j = -\rho_{Y_j Z}$$

as is obvious from equation (1). If we reduce the vector Y to one element we get the particular index computed by Sitgreaves.

It should be noted that a more general set of items than a Guttman scale can be used as long as the vector response is in a finite set. But then the characteristic vector of N will not necessarily depend only on the distribution of the items. The ν_k's are the crucial quantities either scalar or vector-valued, to be computed.

With n dichotomous items there are 2^n possible responses. Once more let us assume a structure examined by Sitgreaves, in that context. (Chap. 3 in Solomon [1961]). We have $X_i = 1$ if a certain random variable Z_i is greater than α_i and equals zero, otherwise. Suppose the Z_i's and the Y_j's are jointly normally distributed. Now the P_k values and the ν_k vectors can be computed explicitly, involving no more (and no less!) than the evaluation of multiple integrals of the multinormal density. Namely:

$$\nu_{kj} = \frac{\int_{S^*} y_j \, \varphi_{n+1}(Z_1, \ldots, Z_n, y_j) \, dy_j \prod_i dZ_i}{\int_S \varphi_n(Z_1, \ldots, Z_n) \prod_i dZ_i}$$

where

$S = \{(Z_1, \ldots, Z_n): \alpha_i \le Z_i, \text{ for } i \in M_k, \alpha_i > Z_i \text{ for } i \notin M_k\}$

$S^* = \{(Z_1, \ldots, Z_n, y_j): (Z_1, \ldots, Z_n) \in S \text{ and } y_j \text{ arbitrary}\}$

M_k: are as follows. Out of 2^n possible vector responses the k^{th} one has a subset of the Z_i's for which $Z_i \ge \alpha_i$ $(X_i = 1)$, the rest of the X_i being 0. The set of i's for which $X_i = 1$ is M_k.

φ_n: the multinormal density with proper covariance matrix

and

$$P_k = \int_S \varphi_n(Z_1, \ldots, Z_n) \prod_i dZ_i \; .$$

But in practical situations one will rather compute the quantities $\hat{\nu}_{kj}$ and $\hat{P}_k$ the sample population equivalent of ν_{ji} and P_k, if one is not interested in test design but in prediction given some data. For instance the sample might look as follows:

$X_{(k)}$	X_1	X_2	Y_1	Y_2
$X_{(1)}$	0	0	2,3,4,5	3,4,8,9
$X_{(2)}$	0	1	2,4,5	6,7,8
$X_{(3)}$	1	0	5,7	5,6
$X_{(4)}$	1	1	9,	4,

or (X_1, X_2, Y_1, Y_2) gives samples $(0, 0, 2, 3)$, $(0, 0, 3, 4)$, $(0, 0, 4, 8)$, $(0, 0, 5, 9)$ from which one easily computes the $\hat{\nu}_k$ and the $\hat{P}_k$ [1]:

Example:

$$\hat{\nu}_{11} = E(Y_1 \; X_{(1)}) = \frac{7}{2}$$

$$\hat{P}_1 = \frac{2}{5}$$

This could be applied to studies like the one used by Solomon in Chapter 11 of Solomon [1961], where four dichotomous items about

---

[1] Note $(X_1^{(1)}, X_2^{(1)}) = (0,0) = X_{(1)}$.

"attitudes toward science" are compared to a "brief IQ test". If one assumes the IQ test is composed of some items, instead of reducing them by a score prior to analysis, one might keep all the information given by the items and attempt simultaneous prediction.

2. Non-Linear PC of a Set of Items: Non-Linear Summarization Assuming No Known Latent Structure.

If, in the preceding discussion, we assume the Y_j's are the X_i s themselves, we have a problem similar to Principal Components. In other words, we are attempting to get an equivalent predictor in a non-linear way. Here

$$P_k \nu_k' \nu_\ell = n_{\ell k} = P_k X_{(k)}' X_{(\ell)}$$

since

$$E[X_i | X_1, \ldots, X_i, \ldots, X_n] = X_i \; .$$

We must first reduce the items by their means since the previous work assumes the predicted variables have mean zero. In the case of dichotomous items, X_i becomes

$$X_i = \begin{cases} 1 - \sum\limits_{k \in S_i} P_k > 0 \; , & \text{if original item is } 1 \\ \\ - \sum\limits_{k \in S_i} P_k < 0 \; , & \text{if original item is } 0 \end{cases}$$

where

$$S_i = \{k: \; X_i^{(k)} = 1\} \; .$$

If, moreover, we have a Guttman scale

$$S_i = \{i, i + 1, \ldots, m\}$$

and

$$X_i = \begin{cases} 1 - \sum_{k=i}^{m} P_k , & \text{if original item is } 1 \\ - \sum_{k=i}^{m} P_k , & \text{if original item is } 0 \end{cases}$$

In the case of a Guttman scale it is clear that the function $f(X)$ defined as $f(X_{(k)}) = k = \sum_i X_i^{(k)}$ gives all the information about each item X_i. But it does not necessarily provide a score, which is, in general, a function of that sufficient statistic. This point is illustrated in the case where we have a latent structure or an underlying variable to be predicted. If we do not have such a structure available, we might still desire to give to each $X_{(k)}$ a score different from k, on the ground that

$$\Pr\{f(X) \geq k\} ,$$

may not be a linear function of k. For instance, if we want to find the relative position of individuals on a real-valued continuum, we are interested in those probabilities. They are, of course, $\sum_{\ell=k}^{m} P_\ell = P_k^*$ where the P_k's are the quantities defined before as $\Pr\{X_{(k)}\}$. $P_k^* = f^*(k)$ is then a possible score.

Another way to use those probabilities is precisely the non-linear PC previously established as the first latent vector of N. We have seen that N involves the quantities $\sum_{\ell=k}^{m} P_\ell$ apart from being evolved from a reasonable criterion.

APPENDIX A

In all that follows R: $n \times n$, is positive definite and T is $m \times n$.

Lemma 1: $R^{-1} T' T$ has the same non-zero roots as $TR^{-1} T'$ which is positive semi-definite, (since it is a covariance matrix). Moreover, to the set of orthogonal characteristic vectors associated with the roots of $TR^{-1} T'$ $\{a_i^*\}$ corresponds the set $\{R^{-1} T' a_i^*\} = \{a_i\}$ associated with the roots of $R^{-1} T' T$ (since there is symmetry, to each root correspond as many orthogonal vectors as indicated by its multiplicity).

Proof: Let 0 designate in the following matrices the sub-matrix with zero entries, of appropriate dimension. Assume $n < m$ so that all the following matrices are $m \times m$. Consider

$$\begin{pmatrix} T & 0 \end{pmatrix} \begin{pmatrix} R^{-1} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} T' \\ 0 \end{pmatrix} = TR^{-1} T' = W$$

and

$$\begin{pmatrix} R^{-1} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} T' \\ 0 \end{pmatrix} (T \quad 0) = \begin{pmatrix} R^{-1} T' T & 0 \\ 0 & 0 \end{pmatrix} = V_o .$$

Using the fact that the roots of AB are the roots of BA, the roots of W are the roots of V_o. But obviously the non-zero roots of V_o are the non-zero roots of $R^{-1} T' T$. Use a similar argument for the case where $n > m$. Let us now prove the second part. We have

$$R^{-1} T' T(R^{-1} T' a_i^*) = \lambda_i(R^{-1} T' a_i^*)$$

if λ_i is the i^{th} root of $R^{-1} T' T$ and $TR^{-1} T'$. This completes the proof of the lemma.

Lemma 2: The function $\lambda(a) = \frac{a' T' Ta}{a' Ra}$ attains an absolute maximum on the domain $R_n - \{0\}$. (n-dimensional Euclidean space excluding the origin.)

Proof: The range of λ, $\lambda(R_n - \{0\})$ is the image by λ of the boundary of the unit sphere, $E. = \{a: \sum_i a_i^2 = 1\}$, as it is obvious from the definition of λ, that $\lambda(ka) = \lambda(a)$. But the boundary of the unit sphere is a compact set and λ is continuous on that set. Therefore $\lambda(E)$ is compact and the lemma follows.

Proof of Theorem 1.1: The quantity to be maximized is

$$\sum_j \mathrm{Var}(Y_j)\, \rho^2_{Y_j S_1} = \frac{a_1' T' T a_1}{a_1' R a_1} = \lambda(a_1)$$

since S_1 is defined as a linear combination $a_1' X$ of the X_i's.

By Lemma 2 we know $\lambda(a_1)$ attains an absolute maximum and a necessary condition for that, is that the partial derivatives with respect to the a_{1i}'s be equal to zero. Let us first write down a property for derivatives of a ratio $\frac{u}{v}$. If $(\frac{u}{v})' = 0$, then

$$u' = \frac{u}{v} v' \tag{1}$$

where the "prime" indicates differentiation and v is assumed always different from zero, which is the case of $a_1' R a_1$ on $R - \{0\}$. Now applying (1) we get

$$\frac{\partial}{\partial a_{1i}} (a_1' T' Ta_1) = \lambda(a_1) \frac{\partial}{\partial a_{1i}} (a_1' Ra_1)$$

for all a_{1i}'s, a set of equations equivalent to:

$$T'Ta_1 = \lambda(a_1)Ra_1$$

or

$$R^{-1} T' T a_1 = \lambda(a_1)a_1 \quad (2)$$

It is easy to verify that if a_1 is a characteristic vector of $R^{-1} T' T$, and λ_1 is the corresponding root, (2) is satisfied, and if (2) is satisfied then a_1 and $\lambda(a_1)$ are such quantities. Therefore, choose the largest root λ_1 of $R^{-1} T' T$ and the corresponding vector a_1. Then

$$\lambda(a_1) = \lambda_1$$

the largest value it can attain. Now by Lemma 1 we know the characteristic vectors of $TR^{-1}T'$, a_i^* are such that $R^{-1} T' a_i^* = a_i$ the characteristic vectors of $R^{-1} T' T$. Normalize the a_i^* such that

$$\sum_j a_{ij}^{*2} = \frac{1}{\lambda_i}$$

where λ_i is a root of either $R^{-1}T'T$ or $TR^{-1}T'$. Then

$$\begin{aligned} a_i' R a_i &= (R^{-1}T'a_i^*)' R(R^{-1}T'a_i^*) \\ &= a_i^{*\prime}TR^{-1}R R^{-1}T'a_i^* = 1 \end{aligned}$$

or the variance of S_1 is 1 if

$$\sum_j a^{*2}_{1j} = \frac{1}{\lambda_1} .$$

Then obviously the vector b_1 of the regression coefficients of the Y_j's on S_1 is

$$b_1 = Ta_1$$

and

$$\sum_j b^2_{1j} = \lambda_1$$

since

$$b_1 = Ta_1 = TR^{-1}T'a^*_1 = \lambda_1 a^*_1 .$$

Note that b_1 is essentially a^*_1 .

Proof of Theorem 2: Let $Y^{(2)}_j$ be $Y^{(1)}_j$ reduced by its regression on S_1, (recall $Y_j = Y_j^{(1)}$)

$$Y^{(2)}_j = Y_j - t^{(1)}_j a_1 S_1$$

where $t^{(1)}_j$ is the j^{th} row of T_1. But

$$S_1 = a_1'X = (R^{-1} T_1' a^*_1)' X$$

and

$$Y^{(2)}_j = Y_j - t^{(1)}_j R^{-1} T_1' a^*_1 a^{*\prime}_1 T_1 R^{-1} X$$

$$= Y_j - \lambda_1 a^*_{1j} a^{*\prime}_1 T_1 R^{-1} X$$

and

$$\text{Cov}(Y^{(2)}_j , X_i) = t^{(1)}_{ji} - \lambda_1 a^*_{1j} a^{*\prime}_1 T_1 R^{-1} r_i$$

where r_i is the i^{th} column of R

$$= t^{(1)}_{ji} - \lambda_1 a^*_{1j} a^{*\prime}_1 t^{(1)}_i$$

where $t_i^{(1)}$ is the i^{th} column of T_1, [1] since $(R^{-1} r_i)' = (0, \ldots, 0, 1, 0, \ldots, 0)$ with 1 in the i^{th} position. Therefore

$$T_2 = T_1 - \{(\lambda_1 a^*_{1j})(a^{*\prime}_1 t_i^{(1)})\}$$
$$= T_1 - \lambda_1 a^*_1 a^{*\prime}_1 T_1$$

and

$$T_2 R^{-1} T_2' = (T_1 - \lambda_1 a^*_1 a^{*\prime}_1 T_1) R^{-1} (T_1 - \lambda_1 a^*_1 a^{*\prime}_1 T_1)'$$
$$= T_1 R^{-1} T_1 - \lambda_1 (\sqrt{\lambda_1} a^*_1)(\sqrt{\lambda_1} a^{*\prime}_1)$$

if we note $a^{*\prime}_1 T_1 R^{-1} T_1' = \lambda_1 a^{*\prime}_1$ in the expansion of the right hand side. It follows that $T_2 R^{-1} T_2'$ is $T_1 R^{-1} T_1'$ reduced by its first root not necessarily distinct from the second one and a corresponding vector of norm 1. Therefore, $T_2 R^{-1} T_2'$ has the positive roots $\lambda_2, \ldots, \lambda_p$ in that order of magnitude and the corresponding vectors $a^*_2, \ldots, a^*_p$, by a proof in Faddeeva [1959] p. 225-6. Hence from Lemma 1:

$$S_2 = a_2' X,$$

where $a_2 = R^{-1} T' a_2'$ the vector associated with the second root λ_2 of $R^{-1} T' T$. The rest of the theorem follows easily except for the assertion that

$$p = \text{row-rank } (T)$$

---

[1] There is here a certain abuse of notation as compared to previous use of $t_j^{(1)}$.

which we prove now. We want to find the rank of $R^{-1} T' T$ or of $TR^{-1} T'$. Assume $m > n$, and that:

$$TR^{-1} T'c = 0$$

Then:

$$c'TR^{-1} T'c = 0 ,$$

and the problem is to find the dimensionality of the null space consisting of such vectors c. But by the positivity of R^{-1}, we have

$$T'c = 0 .$$

Also

$$\begin{pmatrix} T' \\ 0 \end{pmatrix} c = 0 .$$

Therefore c is in the null space of $\begin{pmatrix} T' \\ 0 \end{pmatrix}$ [1] which obviously has dimension $m-[\text{row-rank } (T)]$ which in turn implies rank $(TR^{-1} T')$ is row-rank (T) since:

$$\text{rank}(TR^{-1} T') = m - (\text{dimension of null space of } TR^{-1} T').$$

For $m < n$ a similar argument can be used with the vector c or $\begin{pmatrix} c \\ 0 \end{pmatrix}$.

The set of vectors of regression coefficients $\{b_i\} = \{Ta_i\} = \{TR^{-1} T' a_i^*\} = \{\lambda_i a^*\}$ is such that

---

[1] The converse is trivial: if c is in the null space of $\begin{pmatrix} T' \\ 0 \end{pmatrix}$, it is in the null space of $TR^{-1} T'$.

$$b_i' \, b_j = \delta_{ij} \quad \text{[Kronecker } \delta\text{]}$$

since

$$\lambda_i \, a_i^{*\prime} \, \lambda_j \, a_j^* = \delta_{ij}$$

by orthogonality of the characteristic vectors of $TR^{-1} \, T'$ (which can be chosen orthogonal even if a root has multiplicity greater than one by Faddeeva [1959], p. 45). We already know:

$$\sum_j b_{ij}^2 = \lambda_i$$

if the a_i^* are such that

$$\sum_j a_{ij}^{*2} = \frac{1}{\lambda_i} \, .$$

Finally,

$$\begin{aligned} \operatorname{Cov}(S_i, S_j) &= (a_i^{*\prime} \, TR^{-1}) \, R(R^{-1} \, T' a_j^*) \\ &= \lambda_j \, a_i^{*\prime} \, a_j^* \\ &= 0 \, . \end{aligned}$$

APPENDIX B

Suppose in the studies by Kahl and Davis [1956] and French [1957] we decide on the use of 100 runs to determine for each study the selection of the best $n = 2, 3, 4, \ldots, 10$ items by the random sampling technique developed in Section II. For each of the 100 random selections of n items the trace, the eigenvalue, and the eigenvector connected with the BSLP procedure have been computed and listed. They are not all reproduced here for obvious reasons but a master listing is available at the Statistics Department at Stanford or from the author.

Two listings are reproduced here, one for each study. In each case the situation of interest is the selection of three items from the 100 random choices by the trace or eigenvalue index. The choice in each case appears between two lines drawn on the page and these are the items previously listed on p. 38 in Section II.

The reader who is interested in the computation work developed in Section II is referred to the Program for the computation of the BSLP criterion and other relevant quantities in random item selection. This is written in FORTRAN compatible language for the CDC 1604 Computer.[1] The total time required for 900 runs (from 2 to 10 items) for the 41 x 41 source matrix was approximately 15 minutes. Thus the kind of work developed here is feasible with modern computers.

---

[1] Information can be obtained from the author of the program:

Gilles Cantin
U. S. Navy Postgraduate School
Monterey, California

Kahl-Davis Data

```
100  RUNS WITH       3  ITEMS
NUMBER OF SINGULAR MATRICES=     0
NUMBER OF EIGENVECTORS REJECTED=     0

TRACE      EIGENVALUE                                 EIGENVECTOR                                           ITEMS SELECTED
  9.83103    7.56858     1.000    .503   1.297   .000   .000   .000   .000   .000   .000   .000    111410 0 0 0 0 0 0 0
 10.29888    8.34225     1.000    .822    .961   .000   .000   .000   .000   .000   .000   .000    1015 3 0 0 0 0 0 0 0
  9.47065    7.33895     1.000   1.720    .508   .000   .000   .000   .000   .000   .000   .000    16 719 0 0 0 0 0 0 0
  9.91723    8.50530     1.000    .504    .287   .000   .000   .000   .000   .000   .000   .000     2 410 0 0 0 0 0 0 0
 10.83830    8.96886     1.000    .586   1.276   .000   .000   .000   .000   .000   .000   .000     516 4 0 0 0 0 0 0 0
 10.37362    8.37223     1.000    .853    .903   .000   .000   .000   .000   .000   .000   .000     51110 0 0 0 0 0 0 0
 11.37210    8.59028     1.000   2.066   1.040   .000   .000   .000   .000   .000   .000   .000    14 415 0 0 0 0 0 0 0
  8.69940    7.09798     1.000   1.190   1.549   .000   .000   .000   .000   .000   .000   .000    1613 9 0 0 0 0 0 0 0
  9.97545    8.05002     1.000    .896   1.235   .000   .000   .000   .000   .000   .000   .000    1617 5 0 0 0 0 0 0 0
  9.99765    7.77804     1.000   1.619    .727   .000   .000   .000   .000   .000   .000   .000    121011 0 0 0 0 0 0 0
 10.84771    8.88860     1.000    .000   1.952   .000   .000   .000   .000   .000   .000   .000     9 5 1 0 0 0 0 0 0 0
  9.52970    8.02147     1.000   1.154   1.957   .000   .000   .000   .000   .000   .000   .000    18 8 3 0 0 0 0 0 0 0
 10.45962    8.44203     1.000   -.930  -3.537   .000   .000   .000   .000   .000   .000   .000    1718 2 0 0 0 0 0 0 0
  9.99376    8.60757     1.000   1.519   1.742   .000   .000   .000   .000   .000   .000   .000     3 4 2 0 0 0 0 0 0 0
  9.62483    8.04201     1.000   1.464   1.490   .000   .000   .000   .000   .000   .000   .000    17 7 8 0 0 0 0 0 0 0
 10.49961    8.60235     1.000    .812    .685   .000   .000   .000   .000   .000   .000   .000     3 514 0 0 0 0 0 0 0
 11.01475    8.40028     1.000    .502    .268   .000   .000   .000   .000   .000   .000   .000     21419 0 0 0 0 0 0 0
 10.18037    8.19943     1.000    .619   1.109   .000   .000   .000   .000   .000   .000   .000    1510 4 0 0 0 0 0 0 0
 10.90070    8.50894     1.000   1.394   1.097   .000   .000   .000   .000   .000   .000   .000    10 312 0 0 0 0 0 0 0
  9.61741    7.91033     1.000    .586   1.211   .000   .000   .000   .000   .000   .000   .000    1017 5 0 0 0 0 0 0 0
  9.25307    8.04356     1.000    .993    .983   .000   .000   .000   .000   .000   .000   .000     7 8 5 0 0 0 0 0 0 0
  8.63388    6.54652     1.000   1.469   1.100   .000   .000   .000   .000   .000   .000   .000    161218 0 0 0 0 0 0 0
  9.63544    8.14732     1.000   4.337  -1.214   .000   .000   .000   .000   .000   .000   .000    10 1 4 0 0 0 0 0 0 0
  9.77306    8.06949     1.000   2.448   1.715   .000   .000   .000   .000   .000   .000   .000    17 313 0 0 0 0 0 0 0
 11.55808    8.87999     1.000   1.401   1.720   .000   .000   .000   .000   .000   .000   .000    12 4 2 0 0 0 0 0 0 0
 10.30053    8.31503     1.000   1.531    .954   .000   .000   .000   .000   .000   .000   .000    10 316 0 0 0 0 0 0 0
  9.59306    8.30487     1.000    .601   2.688   .000   .000   .000   .000   .000   .000   .000     7 5 1 0 0 0 0 0 0 0
 10.13386    8.40770     1.000    .453    .519   .000   .000   .000   .000   .000   .000   .000     215 3 0 0 0 0 0 0 0
 10.06398    8.16174     1.000  -1.342  -4.416   .000   .000   .000   .000   .000   .000   .000    17 7 2 0 0 0 0 0 0 0
 10.40544    8.52023     1.000    .469    .515   .000   .000   .000   .000   .000   .000   .000     2 5 9 0 0 0 0 0 0 0
 10.49977    8.45808     1.000    .880    .971   .000   .000   .000   .000   .000   .000   .000     3 812 0 0 0 0 0 0 0
 10.02727    8.00186     1.000   3.008    .331   .000   .000   .000   .000   .000   .000   .000    13 219 0 0 0 0 0 0 0
 10.48913    8.65972     1.000    .225    .810   .000   .000   .000   .000   .000   .000   .000     3 9 6 0 0 0 0 0 0 0
  9.96143    7.78641     1.000   1.291   1.829   .000   .000   .000   .000   .000   .000   .000    1711 7 0 0 0 0 0 0 0
 10.03476    8.19763     1.000    .581    .504   .000   .000   .000   .000   .000   .000   .000     416 8 0 0 0 0 0 0 0
 11.11843    8.51742     1.000   1.607   1.207   .000   .000   .000   .000   .000   .000   .000    13 411 0 0 0 0 0 0 0
 10.95678    8.60866     1.000    .839    .670   .000   .000   .000   .000   .000   .000   .000    101516 0 0 0 0 0 0 0
  9.78326    7.57772     1.000   1.018   1.023   .000   .000   .000   .000   .000   .000   .000    17 916 0 0 0 0 0 0 0
 10.47294    8.63459     1.000    .464    .659   .000   .000   .000   .000   .000   .000   .000     2 8 6 0 0 0 0 0 0 0
 10.08180    7.54320     1.000   1.604   1.942   .000   .000   .000   .000   .000   .000   .000    191712 0 0 0 0 0 0 0
  9.75517    8.55591     1.000    .507   1.995   .000   .000   .000   .000   .000   .000   .000     3 8 1 0 0 0 0 0 0 0
 10.72632    8.30800     1.000    .967    .907   .000   .000   .000   .000   .000   .000   .000    13 712 0 0 0 0 0 0 0
 10.62552    8.51063     1.000    .046   1.432   .000   .000   .000   .000   .000   .000   .000     915 4 0 0 0 0 0 0 0
 11.20943    9.01649     1.000   4.192   1.981   .000   .000   .000   .000   .000   .000   .000    13 1 9 0 0 0 0 0 0 0
 10.97824    9.05157     1.000   1.900   3.741   .000   .000   .000   .000   .000   .000   .000     815 1 0 0 0 0 0 0 0
  8.85651    6.91763     1.000   1.651   2.200   .000   .000   .000   .000   .000   .000   .000    1917 9 0 0 0 0 0 0 0
  8.93758    7.20501     1.000    .358   1.497   .000   .000   .000   .000   .000   .000   .000     919 6 0 0 0 0 0 0 0
  9.39126    7.57901     1.000    .382    .906   .000   .000   .000   .000   .000   .000   .000    1317 9 0 0 0 0 0 0 0
 10.90070    8.50894     1.000    .787    .717   .000   .000   .000   .000   .000   .000   .000     31210 0 0 0 0 0 0 0
 10.61077    7.87178     1.000   -.795   1.437   .000   .000   .000   .000   .000   .000   .000    101412 0 0 0 0 0 0 0
```

Kahl-Davis Data (cont.)

```
  TRACE       EIGENVALUE                          EIGENVECTOR                                               ITEMS  SELECTED

 11.20219     8.29576    1.000    .328   -.179  .000  .000  .000  .000  .000  .000  .000     21217 0 0 0 0 0 0 0
 10.54962     8.70427    1.000    .495    .356  .000  .000  .000  .000  .000  .000  .000     216 8 0 0 0 0 0 0 0
  9.44759     7.48851    1.000   1.693    .471  .000  .000  .000  .000  .000  .000  .000    15 819 0 0 0 0 0 0 0
 10.76038     8.19022    1.000   1.014    .979  .000  .000  .000  .000  .000  .000  .000    1012 6 0 0 0 0 0 0 0
  9.41921     7.47429    1.000   1.194    .358  .000  .000  .000  .000  .000  .000  .000     9 518 0 0 0 0 0 0 0
 11.37210     8.59028    1.000    .961   1.986  .000  .000  .000  .000  .000  .000  .000    1514 4 0 0 0 0 0 0 0
 10.24574     7.46997    1.000   1.849   2.207  .000  .000  .000  .000  .000  .000  .000    181210 0 0 0 0 0 0 0
 11.52511     9.40517    1.000    .650   1.753  .000  .000  .000  .000  .000  .000  .000    1116 2 0 0 0 0 0 0 0
 11.50177     8.99321    1.000   1.103   2.258  .000  .000  .000  .000  .000  .000  .000    1311 1 0 0 0 0 0 0 0
 10.63178     8.63054    1.000    .473    .211  .000  .000  .000  .000  .000  .000  .000     111 5 0 0 0 0 0 0 0
  8.54353     6.66500    1.000    .908    .441  .000  .000  .000  .000  .000  .000  .000     91715 0 0 0 0 0 0 0
 10.18864     8.19845    1.000    .506    .712  .000  .000  .000  .000  .000  .000  .000     4 711 0 0 0 0 0 0 0
 11.84774     9.25778    1.000   2.415   1.321  .000  .000  .000  .000  .000  .000  .000    11 1 9 0 0 0 0 0 0 0
 10.46770     8.47561    1.000   1.051    .509  .000  .000  .000  .000  .000  .000  .000     5 615 0 0 0 0 0 0 0
  9.45074     7.85070    1.000    .871    .597  .000  .000  .000  .000  .000  .000  .000     81018 0 0 0 0 0 0 0
 10.46317     8.22492    1.000    .706    .665  .000  .000  .000  .000  .000  .000  .000     51014 0 0 0 0 0 0 0
 10.09852     8.36489    1.000    .912    .895  .000  .000  .000  .000  .000  .000  .000     8 713 0 0 0 0 0 0 0
 10.27199     7.79816    1.000   2.624   2.465  .000  .000  .000  .000  .000  .000  .000    18 912 0 0 0 0 0 0 0
  9.47153     7.42823    1.000   1.718    .761  .000  .000  .000  .000  .000  .000  .000    14 613 0 0 0 0 0 0 0
  7.85946     6.26793    1.000   2.300   2.791  .000  .000  .000  .000  .000  .000  .000    191613 0 0 0 0 0 0 0
 11.02899     8.50147    1.000   1.144   1.114  .000  .000  .000  .000  .000  .000  .000    1110 9 0 0 0 0 0 0 0
  9.07901     7.29360    1.000    .510    .259  .000  .000  .000  .000  .000  .000  .000     61816 0 0 0 0 0 0 0
  9.98394     8.08150    1.000    .864    .685  .000  .000  .000  .000  .000  .000  .000     51317 0 0 0 0 0 0 0
  8.36501     6.44007    1.000   1.176   1.127  .000  .000  .000  .000  .000  .000  .000    1518 9 0 0 0 0 0 0 0
  9.67475     7.88141    1.000   1.004    .634  .000  .000  .000  .000  .000  .000  .000     81019 0 0 0 0 0 0 0
  9.68528     7.44860    1.000   1.348    .997  .000  .000  .000  .000  .000  .000  .000    181419 0 0 0 0 0 0 0
 10.19585     8.71189    1.000    .691   1.562  .000  .000  .000  .000  .000  .000  .000     2 3 1 0 0 0 0 0 0 0
 10.36681     8.00359    1.000    .701    .595  .000  .000  .000  .000  .000  .000  .000     61215 0 0 0 0 0 0 0
  9.80257     7.71385    1.000   1.227   1.104  .000  .000  .000  .000  .000  .000  .000     6 911 0 0 0 0 0 0 0
 10.71945     8.92072    1.000  10.344  11.228  .000  .000  .000  .000  .000  .000  .000    18 7 6 0 0 0 0 0 0 0
 10.31079     7.83170    1.000    .629   1.340  .000  .000  .000  .000  .000  .000  .000    111910 0 0 0 0 0 0 0
 11.10810     8.90121    1.000   1.400   4.185  .000  .000  .000  .000  .000  .000  .000    1916 1 0 0 0 0 0 0 0
 10.84408     8.06370    1.000   1.244    .542  .000  .000  .000  .000  .000  .000  .000    151014 0 0 0 0 0 0 0
  8.69940     7.09798    1.000    .841   1.302  .000  .000  .000  .000  .000  .000  .000    1316 9 0 0 0 0 0 0 0
  9.13639     7.69773    1.000    .432    .850  .000  .000  .000  .000  .000  .000  .000     818 5 0 0 0 0 0 0 0
  9.43540     7.56390    1.000    .352    .731  .000  .000  .000  .000  .000  .000  .000     819 9 0 0 0 0 0 0 0
 11.26185     8.89598    1.000    .592   1.334  .000  .000  .000  .000  .000  .000  .000     512 4 0 0 0 0 0 0 0
  9.33399     7.36071    1.000   5.980   5.592  .000  .000  .000  .000  .000  .000  .000    191113 0 0 0 0 0 0 0
 10.35365     8.10165    1.000   2.125   1.020  .000  .000  .000  .000  .000  .000  .000    15 318 0 0 0 0 0 0 0
  9.50355     7.31827    1.000   1.026    .628  .000  .000  .000  .000  .000  .000  .000    171618 0 0 0 0 0 0 0
  9.98248     8.12365    1.000  -7.096  -9.209  .000  .000  .000  .000  .000  .000  .000    10 6 4 0 0 0 0 0 0 0
 10.31651     7.98317    1.000    .918    .791  .000  .000  .000  .000  .000  .000  .000     51217 0 0 0 0 0 0 0
 10.94701     8.59496    1.000   2.542    .535  .000  .000  .000  .000  .000  .000  .000    14 118 0 0 0 0 0 0 0
 10.94309     8.92819    1.000196.242  91.580  .000  .000  .000  .000  .000  .000  .000     5 115 0 0 0 0 0 0 0
 10.06586     8.56753    1.000   1.938   2.761  .000  .000  .000  .000  .000  .000  .000     8 4 2 0 0 0 0 0 0 0
 11.03412     8.97538    1.000   1.192   1.626  .000  .000  .000  .000  .000  .000  .000    1510 2 0 0 0 0 0 0 0
  9.75985     7.58242    1.000   1.041    .534  .000  .000  .000  .000  .000  .000  .000     61019 0 0 0 0 0 0 0
 11.48911     8.77467    1.000    .618    .479  .000  .000  .000  .000  .000  .000  .000     2 612 0 0 0 0 0 0 0
  9.75985     7.58242    1.000   1.950   1.873  .000  .000  .000  .000  .000  .000  .000    1910 6 0 0 0 0 0 0 0
  8.54353     6.66500    1.000    .485   1.101  .000  .000  .000  .000  .000  .000  .000    1715 9 0 0 0 0 0 0 0
```

French Data

```
100  RUNS WITH       3  ITEMS
NUMBER OF SINGULAR MATRICES=     0
NUMBER OF EIGENVECTORS REJECTED=     0

TRACE      EIGENVALLE                                   EIGENVECTOR                                         ITEMS  SELECTED

23.85277   20.75624   1.000   1.657   1.704   .000   .000   .000   .000   .000   .000   .000   52531 0 0 0 0 0 0 0
23.78916   19.90529   1.000   1.247   1.909   .000   .000   .000   .000   .000   .000   .000   35 524 0 0 0 0 0 0 0
20.54593   18.29868   1.000    .363    .656   .000   .000   .000   .000   .000   .000   .000   302236 0 0 0 0 0 0 0
19.88722   18.17442   1.000   2.091   1.149   .000   .000   .000   .000   .000   .000   .000   371221 0 0 0 0 0 0 0
20.83670   18.95316   1.000    .993    .643   .000   .000   .000   .000   .000   .000   .000   31 638 0 0 0 0 0 0 0
20.28152   17.46334   1.000    .759   1.412   .000   .000   .000   .000   .000   .000   .000   36 334 0 0 0 0 0 0 0
19.71588   16.70217   1.000    .576    .499   .000   .000   .000   .000   .000   .000   .000   1341 4 0 0 0 0 0 0 0
21.85976   19.02925   1.000   1.713   3.028   .000   .000   .000   .000   .000   .000   .000    82210 0 0 0 0 0 0 0
22.01753   19.17725   1.000    .707   1.602   .000   .000   .000   .000   .000   .000   .000   34 810 0 0 0 0 0 0 0
20.56200   18.02663   1.000   1.080   1.331   .000   .000   .000   .000   .000   .000   .000   232130 0 0 0 0 0 0 0
23.05774   20.18952   1.000   3.898   6.614   .000   .000   .000   .000   .000   .000   .000    73125 0 0 0 0 0 0 0
22.86314   20.18089   1.000    .509   1.260   .000   .000   .000   .000   .000   .000   .000   242133 0 0 0 0 0 0 0
22.59737   18.68937   1.000    .995    .921   .000   .000   .000   .000   .000   .000   .000   282034 0 0 0 0 0 0 0
19.26228   18.00667   1.000   1.298   2.644   .000   .000   .000   .000   .000   .000   .000   351026 0 0 0 0 0 0 0
21.27988   18.24477   1.000    .581    .369   .000   .000   .000   .000   .000   .000   .000   10 237 0 0 0 0 0 0 0
21.97454   19.14668   1.000   1.335   1.092   .000   .000   .000   .000   .000   .000   .000   18 636 0 0 0 0 0 0 0
21.45828   17.34717   1.000   1.008   1.099   .000   .000   .000   .000   .000   .000   .000   38 3 7 0 0 0 0 0 0 0
21.37582   17.94636   1.000    .443    .841   .000   .000   .000   .000   .000   .000   .000    7 436 0 0 0 0 0 0 0
18.93181   16.88516   1.000    .467   1.382   .000   .000   .000   .000   .000   .000   .000   1517 6 0 0 0 0 0 0 0
23.29406   19.80262   1.000   1.460    .426   .000   .000   .000   .000   .000   .000   .000    931 4 0 0 0 0 0 0 0
20.76026   18.89674   1.000    .395    .852   .000   .000   .000   .000   .000   .000   .000   293931 0 0 0 0 0 0 0
20.26932   17.37283   1.000   1.125    .114   .000   .000   .000   .000   .000   .000   .000   22 920 0 0 0 0 0 0 0
24.49353   20.22650   1.000   1.673   1.044   .000   .000   .000   .000   .000   .000   .000    72620 0 0 0 0 0 0 0
21.35379   18.39353   1.000   1.270   1.201   .000   .000   .000   .000   .000   .000   .000   36 915 0 0 0 0 0 0 0
21.27068   19.13234   1.000   3.405   8.067   .000   .000   .000   .000   .000   .000   .000   171033 0 0 0 0 0 0 0
22.53519   20.19020   1.000    .445    .655   .000   .000   .000   .000   .000   .000   .000   263822 0 0 0 0 0 0 0
15.84948   14.12606   1.000   -.341   1.422   .000   .000   .000   .000   .000   .000   .000    2 836 0 0 0 0 0 0 0
21.60941   18.15127   1.000   2.097   2.639   .000   .000   .000   .000   .000   .000   .000    628 5 0 0 0 0 0 0 0
21.71829   18.92912   1.000   -.104   2.695   .000   .000   .000   .000   .000   .000   .000    5 325 0 0 0 0 0 0 0
17.69034   15.49816   1.000    .446    .167   .000   .000   .000   .000   .000   .000   .000    9 420 0 0 0 0 0 0 0
23.94873   20.22435   1.000   3.210   1.306   .000   .000   .000   .000   .000   .000   .000    73314 0 0 0 0 0 0 0
19.21031   16.33623   1.000    .638   -.010   .000   .000   .000   .000   .000   .000   .000   11 2 3 0 0 0 0 0 0 0
23.56421   20.25267   1.000    .826    .493   .000   .000   .000   .000   .000   .000   .000   31 535 0 0 0 0 0 0 0
16.93074   14.56558   1.000    .688    .731   .000   .000   .000   .000   .000   .000   .000   14 3 4 0 0 0 0 0 0 0
21.85461   18.86435   1.000    .694    .763   .000   .000   .000   .000   .000   .000   .000   1018 5 0 0 0 0 0 0 0
21.40456   18.93958   1.000   2.615    .534   .000   .000   .000   .000   .000   .000   .000   162628 0 0 0 0 0 0 0
22.17249   19.02617   1.000   1.963   1.654   .000   .000   .000   .000   .000   .000   .000    340 6 0 0 0 0 0 0 0
21.06836   19.05080   1.000   2.085   1.018   .000   .000   .000   .000   .000   .000   .000   181235 0 0 0 0 0 0 0
19.31376   15.69791   1.000    .731    .274   .000   .000   .000   .000   .000   .000   .000   2314 8 0 0 0 0 0 0 0
20.67418   18.75676   1.000    .553   1.133   .000   .000   .000   .000   .000   .000   .000   351827 0 0 0 0 0 0 0
18.49355   16.28528   1.000   2.410   3.560   .000   .000   .000   .000   .000   .000   .000    82221 0 0 0 0 0 0 0
19.82941   18.39172   1.000    .135    .116   .000   .000   .000   .000   .000   .000   .000   333437 0 0 0 0 0 0 0
22.89129   19.89839   1.000   3.652   1.174   .000   .000   .000   .000   .000   .000   .000   1935 5 0 0 0 0 0 0 0
22.41236   19.51759   1.000   1.963   1.740   .000   .000   .000   .000   .000   .000   .000   161240 0 0 0 0 0 0 0
21.85168   18.94178   1.000   2.631    .894   .000   .000   .000   .000   .000   .000   .000   3626 5 0 0 0 0 0 0 0
21.93636   17.92254   1.000   1.686   1.374   .000   .000   .000   .000   .000   .000   .000   17 228 0 0 0 0 0 0 0
20.44262   18.82944   1.000   7.398   1.031   .000   .000   .000   .000   .000   .000   .000   142519 0 0 0 0 0 0 0
22.48914   19.71354   1.000    .262    .435   .000   .000   .000   .000   .000   .000   .000   33 827 0 0 0 0 0 0 0
19.33175   16.42863   1.000    .910   1.405   .000   .000   .000   .000   .000   .000   .000    22017 0 0 0 0 0 0 0
20.63651   18.94810   1.000    .421    .805   .000   .000   .000   .000   .000   .000   .000   31 140 0 0 0 0 0 0 0
```

French Data (cont.)

```
TRACE      EIGENVALUE                                  EIGENVECTOR                                                ITEMS  SELECTED

19.23143   16.98914   1.000  1.113   .628   .000   .000   .000   .000   .000   .000   .000   361819 0 0 0 0 0 0 0
20.66735   18.26767   1.000   .494   .517   .000   .000   .000   .000   .000   .000   .000   401315 0 0 0 0 0 0 0
20.10339   17.18305   1.000   .669  1.147   .000   .000   .000   .000   .000   .000   .000    41528 0 0 0 0 0 0 0
20.64610   17.82797   1.000   .268  1.098   .000   .000   .000   .000   .000   .000   .000   3619 6 0 0 0 0 0 0 0
22.80218   20.13069   1.000  1.099   .635   .000   .000   .000   .000   .000   .000   .000   2925 5 0 0 0 0 0 0 0
21.87538   18.66936   1.000   .484   .102   .000   .000   .000   .000   .000   .000   .000   2636 3 0 0 0 0 0 0 0
22.71271   19.86246   1.000   .182   .372   .000   .000   .000   .000   .000   .000   .000   331623 0 0 0 0 0 0 0
21.27295   18.52797   1.000   .486   .232   .000   .000   .000   .000   .000   .000   .000   292332 0 0 0 0 0 0 0
22.50404   19.23242   1.000  1.260  1.181   .000   .000   .000   .000   .000   .000   .000   164021 0 0 0 0 0 0 0
23.10452   20.05735   1.000   .592   .390   .000   .000   .000   .000   .000   .000   .000   241220 0 0 0 0 0 0 0
19.87847   17.67219   1.000   .637   .551   .000   .000   .000   .000   .000   .000   .000   113832 0 0 0 0 0 0 0
20.68379   18.52601   1.000   .225   .281   .000   .000   .000   .000   .000   .000   .000   29 739 0 0 0 0 0 0 0
22.49310   19.22042   1.000   .322   .217   .000   .000   .000   .000   .000   .000   .000   252120 0 0 0 0 0 0 0
22.60954   20.36077   1.000   .719   .672   .000   .000   .000   .000   .000   .000   .000   302913 0 0 0 0 0 0 0
22.67659   20.75577   1.000  1.982  1.855   .000   .000   .000   .000   .000   .000   .000   122524 0 0 0 0 0 0 0
22.81517   19.69370   1.000   .817   .910   .000   .000   .000   .000   .000   .000   .000   35 527 0 0 0 0 0 0 0
17.75919   15.37659   1.000   .491   .840   .000   .000   .000   .000   .000   .000   .000   22 1 2 0 0 0 0 0 0 0
20.59239   18.41106   1.000  1.216   .727   .000   .000   .000   .000   .000   .000   .000   221118 0 0 0 0 0 0 0
17.87366   15.54555   1.000   .652   .497   .000   .000   .000   .000   .000   .000   .000   363220 0 0 0 0 0 0 0
19.37515   16.53922   1.000  1.604   .569   .000   .000   .000   .000   .000   .000   .000   1934 8 0 0 0 0 0 0 0
19.85552   17.93548   1.000  1.982  3.425   .000   .000   .000   .000   .000   .000   .000    11810 0 0 0 0 0 0 0
20.70875   18.12794   1.000   .800  2.056   .000   .000   .000   .000   .000   .000   .000   19 910 0 0 0 0 0 0 0
21.01424   18.88155   1.000   .542  1.263   .000   .000   .000   .000   .000   .000   .000   133240 0 0 0 0 0 0 0
22.07239   19.63872   1.000   .861   .452   .000   .000   .000   .000   .000   .000   .000   124017 0 0 0 0 0 0 0
22.10006   18.88459   1.000   .630  1.205   .000   .000   .000   .000   .000   .000   .000   213911 0 0 0 0 0 0 0
20.92735   18.43054   1.000  3.288  2.092   .000   .000   .000   .000   .000   .000   .000   223016 0 0 0 0 0 0 0
19.70585   17.28377   1.000   .833  1.279   .000   .000   .000   .000   .000   .000   .000   152838 0 0 0 0 0 0 0
23.89257   20.33870   1.000  4.343  3.103   .000   .000   .000   .000   .000   .000   .000   202530 0 0 0 0 0 0 0
20.73773   17.94287   1.000   .515   .459   .000   .000   .000   .000   .000   .000   .000   112839 0 0 0 0 0 0 0
22.85632   20.33541   1.000   .391   .701   .000   .000   .000   .000   .000   .000   .000   332824 0 0 0 0 0 0 0
22.47881   19.49383   1.000   .428   .983   .000   .000   .000   .000   .000   .000   .000   211913 0 0 0 0 0 0 0
21.02968   18.61122   1.000  1.676  2.728   .000   .000   .000   .000   .000   .000   .000   393410 0 0 0 0 0 0 0
20.99704   19.19938   1.000   .075  2.409   .000   .000   .000   .000   .000   .000   .000   273733 0 0 0 0 0 0 0
17.35167   15.18369   1.000  1.219  1.038   .000   .000   .000   .000   .000   .000   .000   381817 0 0 0 0 0 0 0
20.54120   18.37479   1.000  1.086   .437   .000   .000   .000   .000   .000   .000   .000   214037 0 0 0 0 0 0 0
22.28384   19.21847   1.000   .849   .383   .000   .000   .000   .000   .000   .000   .000   3040 3 0 0 0 0 0 0 0
21.43434   18.80660   1.000   .709  1.252   .000   .000   .000   .000   .000   .000   .000    6 212 0 0 0 0 0 0 0
22.34313   20.03782   1.000  1.114   .354   .000   .000   .000   .000   .000   .000   .000   253313 0 0 0 0 0 0 0
21.11244   19.09381   1.000  2.727   .720   .000   .000   .000   .000   .000   .000   .000   182437 0 0 0 0 0 0 0
19.86086   16.91670   1.000  1.360  1.046   .000   .000   .000   .000   .000   .000   .000   1513 3 0 0 0 0 0 0 0
21.23263   17.45660   1.000   .734  1.261   .000   .000   .000   .000   .000   .000   .000   202223 0 0 0 0 0 0 0
20.49803   18.42730   1.000   .986   .447   .000   .000   .000   .000   .000   .000   .000   402141 0 0 0 0 0 0 0
23.36162   19.43762   1.000  1.496   .936   .000   .000   .000   .000   .000   .000   .000   202928 0 0 0 0 0 0 0
20.86777   17.65007   1.000   .736   .880   .000   .000   .000   .000   .000   .000   .000   341916 0 0 0 0 0 0 0
20.70129   18.14312   1.000  3.648  1.875   .000   .000   .000   .000   .000   .000   .000   391016 0 0 0 0 0 0 0
23.08155   19.77680   1.000   .655  1.018   .000   .000   .000   .000   .000   .000   .000   402010 0 0 0 0 0 0 0
23.09899   20.43407   1.000   .740   .433   .000   .000   .000   .000   .000   .000   .000   24 940 0 0 0 0 0 0 0
17.89748   15.62611   1.000   .900   .122   .000   .000   .000   .000   .000   .000   .000   222832 0 0 0 0 0 0 0
21.18477   18.48227   1.000   .557   .658   .000   .000   .000   .000   .000   .000   .000    53835 0 0 0 0 0 0 0
19.94080   17.72886   1.000   .657  1.343   .000   .000   .000   .000   .000   .000   .000   343721 0 0 0 0 0 0 0
```

NOTATION

R, S, T, U	are covariance matrices.
X, Y, S	are random vectors.
X_i, Y_j, S	are random real-valued variables.
a_i^*, a_i, b_j	are fixed vectors.
a_{ij}^*, a_{ij}, b_{jk}	are fixed scalars.
t_i, r_i	are row-vectors of T, R.
$\{X_i\}, \{Y_j\}$	are sets of random variables.
T'	is the transpose of T.
BSLP:	abbreviation for Best Simultaneous Linear Predictor.
PC:	abbreviation for Principal Component.
λ_1:	largest or "first" root of a matrix.
λ_i:	i^{th} root (in order of magnitude) of a matrix.

BIBLIOGRAPHY

[1] Anderson, R. L. and Bancroft, T. A., *Statistical Theory in Research*, McGraw-Hill, New York, 1952.

[2] Anderson, T. W., *Introduction to Multivariate Statistical Analysis*, John Wiley and Sons, New York. Second Edition, July, 1960.

[3] Blackwell, David and Girschick, M. A., *Theory of Games and Statistical Decisions*, John Wiley and Sons, New York, 1954.

[4] Faddeeva, V. N., *Computational Methods of Linear Algebra*, Dover Press, New York, 1959.

[5] French, Grace M., "College Students' Concept of Effective Teaching Determined by an Analysis of Teacher Ratings," Mimeographed Study, 1957, University of Washington, Seattle, Washington.

[6] Hotelling, Harold, "Analysis of a Complex of Statistical Variables into Principal Components," *The Journal of Educational Psychology*, 1933, **24**, 417-441 and 498-520.

[7] _____, "The most predictable criterion," The Journal of *Educational Psychology*, 1935, **26**, 139-142.

[8] _____, "Relations between two sets of variables," *Biometrika*, **28**, 1936, 321-377.

[9] Kahl, Joseph A. and Davis, James H., "A Comparison of Indexes of Socio-Economic Status," *American Soc. Review*, 1956, **21**, 317-32.

[10] Solomon, Herbert, *Studies in Item Analysis and Prediction*, Stanford University Press, Stanford California, 1961.

CLUSTERING PROCEDURES

by

J.J. Fortier and H. Solomon

Introduction and Summary.

It is often useful to group individuals, or test items, or tests, or any set of elements, in accordance with the value obtained for a measure of resemblance or a measure of proximity. This measure is a function of the outcomes for a set of variables and provides a unique mapping for each set of responses (uni- or multi-dimensional) into one of several classes. This type of problem arises in many ways and is a familiar one in science whenever an operational summarization of data is attempted to assign individuals into groups.

Educators, psychologists, and others have a pragmatic interest in these grouping procedures and the difficulties encountered are referred to in a global sense as the clustering problem. An intuitive criterion for assessing clustering procedures is one which operates such that the groups or the clusters which are formed are homogeneous "within" a cluster (elements in a cluster are similar) and heterogeneous "between" clusters (elements from different clusters are not alike). Naturally the groups which are produced by a clustering criterion should always be examined for substantive meaning. The difficulties stressed in analyses heretofore have related to 1) the establishment of measures of homogeneity and heterogeneity and 2) the construction of clustering procedures which minimize the relative frequencies of the two types of misclassification that are possible. Still another difficulty which usually follows some success in these two problems is the management of the excessive enumeration which occurs in many studies.

Therefore, an assessment of a proposed clustering procedure requires first an evaluation of two measures of error: 1) The error incurred by grouping dissimilar elements in the same cluster. 2) The error incurred by not grouping similar elements in the same cluster. Measurement of this error leads to the examination of general loss functions, which for a given structure of clusters, give a value to any pair of variables whether clustered or not; so that a resultant value obtained by summing over all such pairs is the quantity by which we may judge simultaneoulsy the worth of both the number of clusters produced and their configuration.

The new clustering criterion developed in this paper which satisfies this type of risk function gives a precise method for the construction of clusters and minimizes the risk which is measured by this loss function. This precise method can lead swiftly to almost insurmountable computing problems even when large scale electronic digital computers are available. Approximations are required to make an analysis somewhat manageable. Probability sampling provides one approach which reduces the computing load and yet yields results which are capable of quantitative assessment. Simple random sampling is discussed and explored in this paper. Other forms of probability sampling may be more appropriate and they require further study.

A more general version of the same problem is sketched in the last section where n-tuplets instead of pairs of variables (2-tuplets) become the basic elements for the clustering analysis. In that situation the "factor structure" of the variables under study is employed to help understand the goals and methods of clustering.

1. The B clustering function: an unsuccessful attempt.

In what follows we use the correlation coefficient ρ as a measure of distance between two variables. By an orthogonal linear transformation the variables can be located on the unit sphere in a space of orthogonal factors obtained through the Principal Components method, then ρ is a monotonic function of the Cartesian distance and, like distance, ρ is invariant under any linear transformation. Actually any measure of distance could be used for the clustering index we now discuss.

The ratio of the average ρ (assume all ρ's are positive) for pairs belonging to the same cluster to the average ρ for pairs not belonging to the same cluster seems, at first, a reasonable function to maximize. It is a version of the value developed by Holzinger [1949] who considered it as a coefficient of "belongingness" and labelled it B. It is similar to the index employed by Tryon [1939], as an index for grouping variables. We observe here that in the "belongingness" procedure the structure (number and configuration) of clusters is given in advance and the criterion is applied as a checking or inspection device. In this way the clusters are initiated directly by some substantive arguments and then modified afterwards by inspection of the B values.

Theoretically speaking, we have a finite problem at hand for which there exists a solution: i) Prepare a list of all possible partitions of a set of n points. ii) For each partition (set of clusters) calculate the B function. iii) Choose the clustering configuration which corresponds to the optimal (large) B value. However, matters can get out of hand rather quickly for the amount of computation becomes inordinately large even when n is moderate in size.

Some previous work in item selection by Fortier [1] used data from an article by Kahl and Davis [1956] to demonstrate some results. The extensive computational analysis applied to their basic data can be used to good advantage if we permit their work to serve as an illustrative example again. The Kahl and Davis data is based on 19 variables which we will now subject to cluster analysis. Suppose, for example, that we are interested in the number of partitions of those 19 variables in, say, 8 clusters and that for each such partition we will compute our version of the B value. We might then do the same for all possible clusters obtained by dividing up the 19 variables. That is, we might do all this if it is feasible and thus an analysis of the combinatorial situation is required. The development which follows leads to the following recursive formula.

$$P(m) = (m^n - \sum_{i=1}^{m-1} m_{(m-i)} P(i))/m!$$

where: m = number of clusters or partitions, $m \geq 2$

n = number of variables in the set , $n \geq m$

$P(m)$ = number of distinct partitions containing exactly m clusters

$m_{(m-i)} = m(m-1)(m-2) \cdots (m-i+1)$.

To prove this, one can see the clusters as m boxes indexed from 1 to m in which the n variables are to be distributed. If none of the boxes is empty the number of ways one can distribute the n variables corresponds to the number of partitions of the first n integers into m categories except for the $m!$ permutations of the indexes. However, from the m^n ways of distributing the variables one must subtract the number of situations where there are exactly $(m-i)$ empty boxes whenever there are $i < m$ clusters. In order to do this recall that a distribution with i non-empty boxes is a partition with i clusters except for the permutation of the indexes. Since there are $\binom{m}{i}$ ways of choosing the non-empty boxes, one must also multiply $P(i)$ by that number. So there are $\binom{m}{i}i!\ P(i)$ partitions, or $m_{(m-i)}\ P(i)$ partitions, containing $(m-i)$ empty boxes. If we sum over i from 1 to $(m-1)$ and subtract from m^n we have the number of partitions with $(m-i)$ non-empty boxes. Dividing finally by $m!$ to allow for permutation of the indexes we arrive at the formula for $P(m)$.

To help answer the type of question raised in the previous paragraphs a table of the formula was computed for $n=1,2,3,\ldots,19$ and for each $n, m=1,2,\ldots,n$. This appears as Table 1 and contains entries correct to eight places. Actually there are 1, 709, 751, 003, 480 distinct partitions of 19 variables in 8 clusters. The Table gives 1, 709, 750, 700,000 as an approximation. To examine them all is out of the question even with the present state of computer technology. Thus the following strategy based on unrestricted random sampling was employed. It was developed originally for an item selection problem and is discussed in [1].

TABLE 1

Number of Distinct Partitions of n Variables Containing Exactly m Clusters

m \ n	3	4	5	6	7	8	9	10	11	12	13	14	......	19*
2	3	7	15	31	63	127	255	511	1023	2047	4095	8191		262143
3		6	25	90	300	966	3025	9330	28501	86525	261625	788970		19344810×10^1
4			10	65	350	1701	7770	34105	145750	611501	2532530	10391745		11259666×10^3
5				15	140	1050	6951	42525	246730	1379400	7508501	40075035		14758928×10^4
6					21	266	2646	22827	179487	1323652	9321313	63436373		69308159×10^4
7						28	462	5880	63987	627396	5715424	49329283		14929248×10^5
8							36	750	11879	159027	1899612	20912321		17097507×10^5
9								45	1155	22275	359502	5135127		11446150×10^5
10									55	1705	39325	752756		47729690×10^4
11										66	2431	66064		12941322×10^4
12											78	3368		23466998×10^3
13												91		28924104×10^2
14														24358702×10^1
15														13916309
16														526029
17														13270
18														171
:														

* Approximate values

From the laws of simple random sampling we can compute how many cluster selections are to be attempted and enumerated before one attains a probability α of having captured one of the selections which rank among the best. The number of selections desired is thus a function of a predetermined β, the proportion of all the selections which can be labelled "best" or acceptable in some sense. In our present setting all "B" values larger than a preassigned constant determine those clusterings which are considered good and these together are a proportion β of the total number of possible clusterings. From this point on, "B" represents our version of Holzinger's B. This function of α and β, say, $s(\alpha,\beta)$, is the number of samplings required and is given by

$$s(\alpha,\beta) = \left[\frac{\log \alpha}{\log(1-\beta)}\right]$$

where the brackets denote the smallest integer greater than or equal to the value enclosed. A table of $s(\alpha,\beta)$ appears in [1] for values of α, β ranging from .0001 to .20 and is reproduced here.

Table of $m = s(\alpha,\beta)$, for certain values of α and β.

$\beta \diagdown \alpha$	.20	.10	.05	.01	.001	.0001
.20	8	11	14	21	31	42
.10	16	22	29	44	66	88
.05	32	45	59	90	135	180
.01	161	230	299	459	689	918
.001	1,626	2,326	3,026	4,652	6,977	9,303
.0001	17,475	25,000	32,526	50,000	75,000	100,000

Using a computer, 10,000 clusterings were obtained by unrestricted random sampling for each fixed m, and to each clustering a computed "B" value was attached. This was done for values of m=2,...,18; (170,000 clusterings in all). The results were disappointing and so will not be discussed here in detail. Indeed the clustering for the data in Kahl and Davis [1956] using Tryon's method, which constructs clusters sequentially and yielded 8 clusters, has a "B" value superior to that "B" value for our best sample clustering for m=8.

This fact seems to indicate that simple random sampling is not effective when the distribution of the index is very skewed and favorable values of the index are in the tail. For example if $\alpha = .05$, then the number of partition samples required jumps quickly from 200 to 3026 to 32,526 as β goes from .01 to .001 to .0001. To get some idea of the actual frequency distribution of maximum values of "B", an empirical distribution for m=10 was obtained as follows. The 10,000 sample partitions were viewed now as 100 random repetitions each producing 100 partitions of 19 variables in 10 clusters. For each partition, a "B" was computed and the maximum of the 100 values so obtained was recorded. Since there are 100 repetitions, we get 100 such maximum B values. These are given in histogram form in Figure 1 and indicate an exponential type distribution. This is not unusual for ratio type variables and for extreme value variables. Thus we may very well be in a situation where for reasonable values of α, β and for unrestricted random sampling there is a horde of either "poor" or "not good" solutions and a minute number of "good" solutions; and extensive sampling, and thus high cost

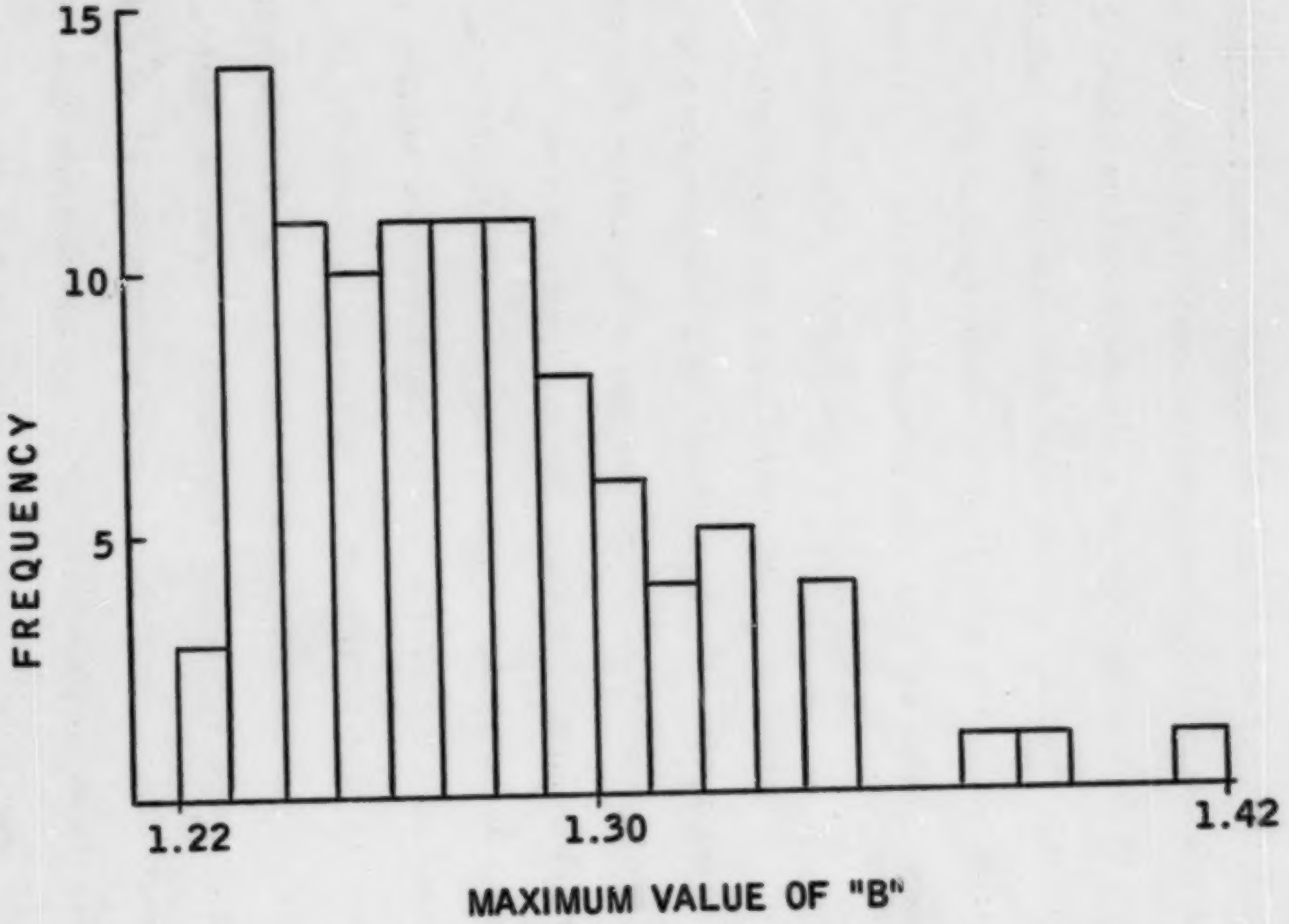


Fig. 1 - Frequency Distribution of Maximum "B" Values for 100 sets of 100 Selections for m = 10 (Distribution of 100 maximum values)

or terrible tedium, represents the only way out. Modifications of the sampling strategy may improve the situation and these should be studied.

2. A clustering function which is defined for each pair of variables.

The previous discussion indicates that we should look for some new clustering indexes if we wish to examine a cluster structure globally (non-sequentially). It would be nice if this new index is more indicative in advance as to whether a certain pair of variables should be assigned to the same cluster or not. Our previous methodology did not employ prior knowledge, it was based on "a posteriori" information arrived at by sampling techniques. We would like a procedure which could lead to an amenable solution by eliminating in advance most of the "poor" solutions. We would also like to give a cumulative form to the clustering function which is defined for each pair, so that the distribution of the clustering index can be approximated, possibly by a Gaussian distribution if one of the Central Limit Theorems applies. This would mitigate the kind of problem we encountered previously if sampling is required since it would permit a value for β from Gaussian tables for any value of the clustering index which is labelled "good".

First we suggest that two variables should be clustered if ρ^2 is greater than a preassigned constant, say ρ^{*2}. In the contrary case they should not be in the same cluster. The gain (positive or negative) incurred by taking any action "a" could be expressed by $G_{ij}(a)$. Suppose we let $\rho^{*2} = .50$ for purposes of exposition. In actual practice, $\rho \geq .7$ signifies a close relationship. Then, we may write

$$G_{ij}(a) = (\rho_{ij}^2 - .5)\cdot g(a)$$

where

$$g(a) = \begin{cases} +1, & \text{if the } i^{th} \text{ and } j^{th} \text{ variable are put in the same cluster} \\ -1, & \text{otherwise.} \end{cases}$$

Summing over all pairs we obtain a value, call it C, for any clustering:

$$C(A) = \sum_{i<j} G_{ij}(a_{ij})$$

where a_{ij} is the specific action taken for the pair (i,j), and A is the matrix of those a_{ij}'s. We observe that some a_{ij}'s depend on others. For instance if the pair (2,3) is in the same cluster, and this is so for the pair (3,4), then the pair (2,4) is also in the same cluster.

Let us now consider the critical values

$$D_{ij} \equiv (\rho_{ij}^2 - .5)$$

The matrix D consists of a number of positive values. For example, the Kahl and Davis correlation matrix leads to a D matrix with 16 positive values. This demonstrates in this case that from one to 16 clusters may be possible since a cluster formed only by pairs having negative D_{ij} values should not be considered since it reduces the gain function (increases the loss function). There may be less than 16 clusters as we will see subsequently even if one chooses to include all 16 pairs. A closer examination of the problem shows that

$$C(A) = 2 \sum_{(i,j \in S)} D_{ij} - \sum_{i<j} D_{ij}$$

where S is the set of pairs of variables that belong to the same cluster. Thus the critical quantity is

$$C^* \equiv \sum_{(i,j \in S)} D_{ij}$$

This is so because $\sum_{i<j} D_{ij}$ is a constant for it is the sum of the elements of the lower half of the D matrix and each element is fixed when ρ_{ij} and ρ^* are given.

Thus we would like C^* to be as large as possible and we hope to achieve this maximum value by a prudent selection of positive D_{ij}'s. Indeed if we group only pairs with negative D_{ij}'s one can do better by not clustering at all. Therefore any reasonable clustering demands at least one pair with positive D_{ij} in the same cluster. We must remember that the selection of other pairs to be grouped in the same cluster, leads implicitly to other D_{ij}'s which must be included in the calculation of C^*. Of course we would like many positive D_{ij}'s after the first positive D_{ij} but each addition can implicitly introduce other D_{ij}'s which may be negative and thus outweigh what is ostensibly a gain. As a matter of fact the choice of any number of D_{ij}'s determines the configuration of the clusters. For example, the choice of D_{12}, D_{23}, D_{45} provides the clusters:

Cluster 1	Cluster 2
Variable 1	Variable 4
Variable 2	Variable 5
Variable 3	

since Variable 1 and Variable 2 are in a cluster, Variable 2 and Variable 3 are in a cluster, and thus Variables 1, 2 and 3 are together in the same cluster. Variable 4 and Variable 5 are grouped in one independent cluster and the other variables are left alone. It is to be noted here that D_{13} is implicitly introduced in the sum C^*. In general the choice of say p of the D_{ij}'s leads to $m \leq p$ clusters and to $q \geq p$ elements in the sum C^*. If $m < p$ then $q > p$, since at least 2 pairs (3 or more variables) are grouped together introducing one or more new D_{ij} values.

In mathematical terms one must maximize C^*, but choosing the appropriate partition of the n variables into m clusters still requires inspection in addition to algebra. In the Kahl and Davis matrix, for instance, one has to choose p pairs out of 16 pairs, decide which D_{ij}'s are also implicitly chosen and then sum these together with the p D_{ij}'s already chosen. This process is repeated for another set of p pairs. There are, of course $\binom{16}{p}$ possible combinations and these represent a heavy load of work for all values of p from 2 to 15, even when using a large and fast computer. This was done for all values of p for the Kahl-Davis data and the maximum C^* obtained for each p. These results are graphed in Figure 2. This demonstrates that the optimal value of p is 6 or 7 which as we shall describe later leads to five clusters $(m=5)$.

If all the ρ's were equal the criterion would lead to one obvious solution: if $\rho^2 > .5$, group all the variables in one cluster since all the D_{ij}'s are positive; otherwise do not cluster at all, since the D_{ij}'s are all negative. This "all or none" behavior does not depend

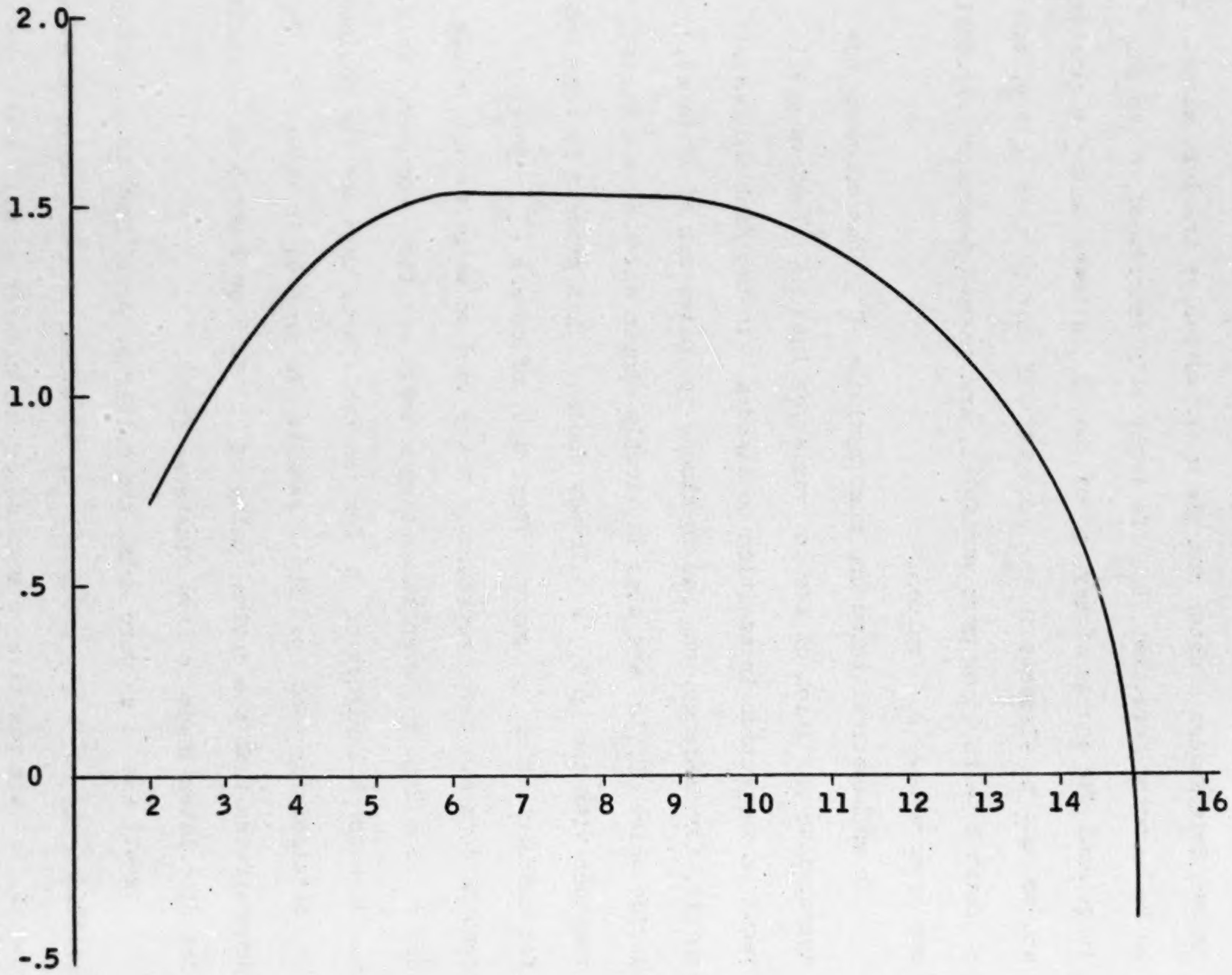


Fig. 2 - Maximum C^* as a Function of p

on the constant .5 and is valid for any preassigned value for ρ. As a matter of fact, if the distance between any pair of variables is constant there is no difference in choosing one configuration over another since it is impossible to decide to which cluster one should assign a variable.

A program has been developed in Fortran language to handle the procedure described above and has been applied to the Kahl-Davis data. Naturally the type of computer available will determine both the size of the correlation matrix and the value of p that can be analyzed. This program involves tricky manipulations in order to account for all the pairs that are implicitly selected through other pairs. However a saving of effort was anticipated by the hope that the maximum C^* for each value of p would increase for a while, then reach a plateau and decrease sharply, because the addition of more positive D_{ij}'s is outweighed by the implicit addition of negative D_{ij}'s. This would mean that one would stop computing after reaching that plateau thus saving much labor. For the Kahl-Davis data, the results behave exactly as heuristically expected. The optimal value of p shown in Figure 2 is 6 or 7, and leads essentially to the clustering obtained by Kahl and Davis using Tryon's method.

The following outline demonstrates the clustering results using the C^* method and Tryon's method on the Kahl-Davis data.

	Tryon's method	C* method
Clusters*	Variables	
1	12. Area rating 14. House rating	The same variables
2	15. Subject's father's education 9. Subject's mother's education	The same variables
3	2. Friend's occupation 17. Wife's education	The same variables
4	4. Subject's occupation Census 1. " " Warner 10. Source of income	The same variables
5	16. Wife's father's occupation North-Hatt 17. Wife's father's occupation	Not a cluster
6	11. Census tract 18. Income	Not a cluster
7	3. Subject's education 8. Subject's self identification	The same variable
8	19. Subject's father's occupation, North-Hatt 13. Subject's father's occupation, Census	Not a cluster

* The clusters are ordered by decreasing values of Tryon's index.

We observe a striking resemblance between the two results though they were arrived at by completely different methods. We could probably arrive at almost the same result by inspection also. But this is not the point, since our purpose is an objective method which is amenable to analysis and does not exhaust the capacity of digital computers.

However, there is an interesting difference in the results obtained by each of the two methods. In the C* result, some clusters are conspicuously absent. Being less homogeneous they may have been eliminated through the choice of $\rho^{*2} = .50$. But the 7^{th} cluster is present and this suggests that changing the value of ρ^* may never produce the same clustering by the two methods. One possible explanation is that Tryon's is a sequential method and the other method (C*) is a global procedure. However, this one difference does not detract from a most extraordinary similarity of clustering results which also have substantive meaning.

Figure 3 gives a distribution for the C* values obtained in the case where $p = 4$. It is roughly normal and this indicates feasibility of sampling. For example, one can estimate the variance of the distribution of C* on the basis of some selections and then proceed to fix a level for the β value under the assumption of normality. With this knowledge, or an assumption of it beforehand, the clustering could have been achieved more cheaply. However, it would be good to get the empirical distribution of C* for other situations when p takes on values other than 4 to see if the normal distribution assumption is still realistic.

3. A generalization of the C* criterion.

In many studies concerned with factorial analysis the problem is not only one of reducing the space generated by the variables under study into a factor space of fewer dimensions, but also that of assigning subsets of variables to subspaces of this factor space.

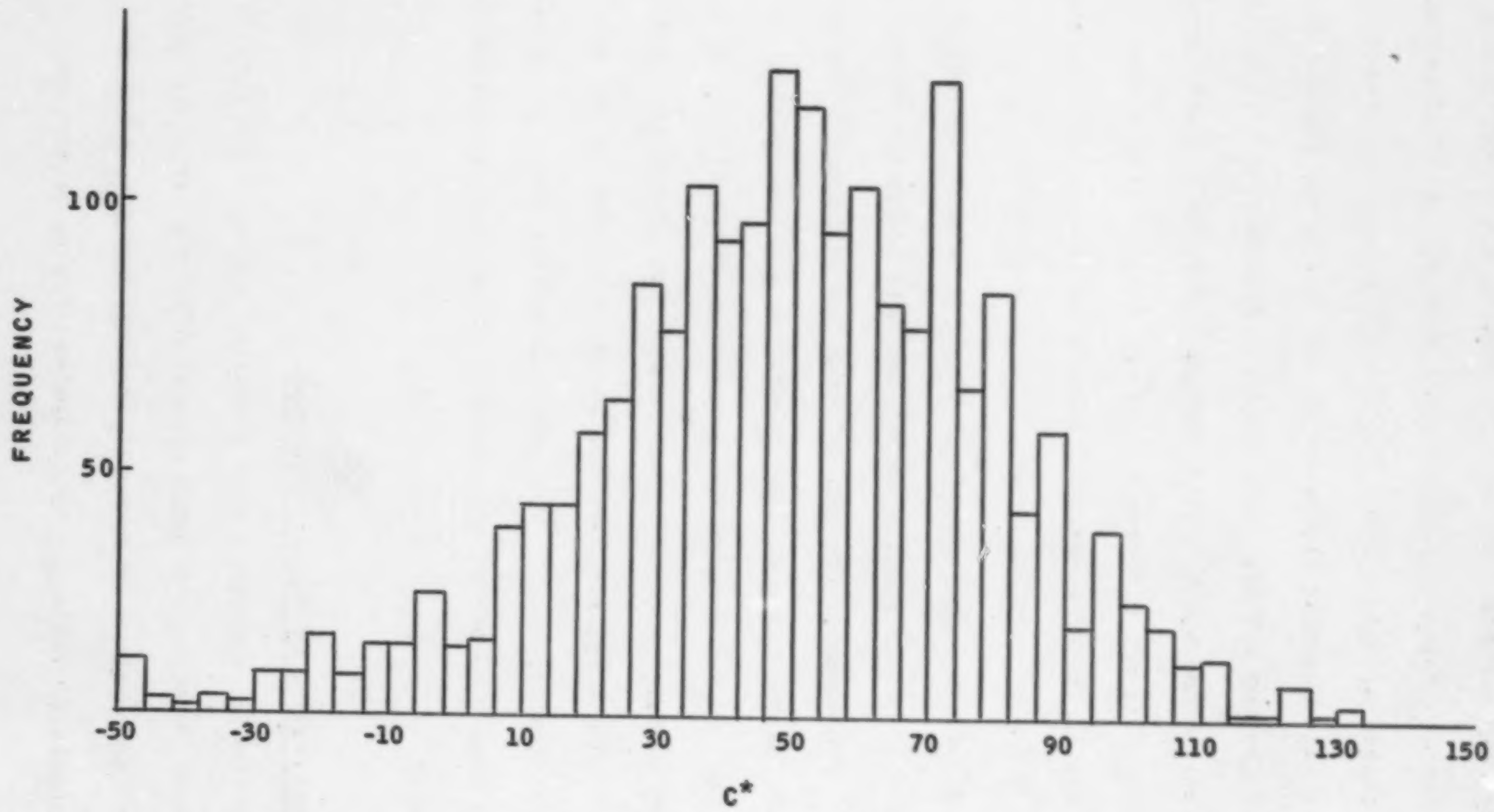


Fig. 3 - Frequency Distribution of C^* for $p = 4$

At least this has been attempted by several contributors to the field, for example through various rotational methods to achieve simple structure.

In a manner similar to those methods, the C* method can be viewed as an attempt to reduce the space of factors to subspaces of one dimension (i.e., our clusters). Indeed if two variables are heavily "loaded" in the same unique factor their correlation will tend to ± 1. This is true for all pairs belonging to the same factor or cluster.

This suggests that we try to view clustering in the framework of a specialized factor structure where each of the subspaces are disjointed and are of dimensionality $k \geq 1$. By this we mean that we will divide the factor space into subspaces of k-dimensions or less if possible. Note that for $k = 1$, the structure will not vary in substance from what is pictured in Figure 4; this is true also for $k = 3$ except that more configurations are possible. In that framework a variable belonging to a cluster must have its variance explained, except for specific factors, by factors which make up the cluster. A general loss function, given sub-factor spaces of dimensionality k would be defined by all $(k+1)$-tuplets of variables of the sub-factor space. The dimensionality can be assessed in a number of ways; for instance by extracting the roots of the covariance matrix (that is, applying the Principal Component method). However this is an indirect way of clustering and we have proposed more direct procedures which unfortunately depend on the structure which is usually not known in the first place.

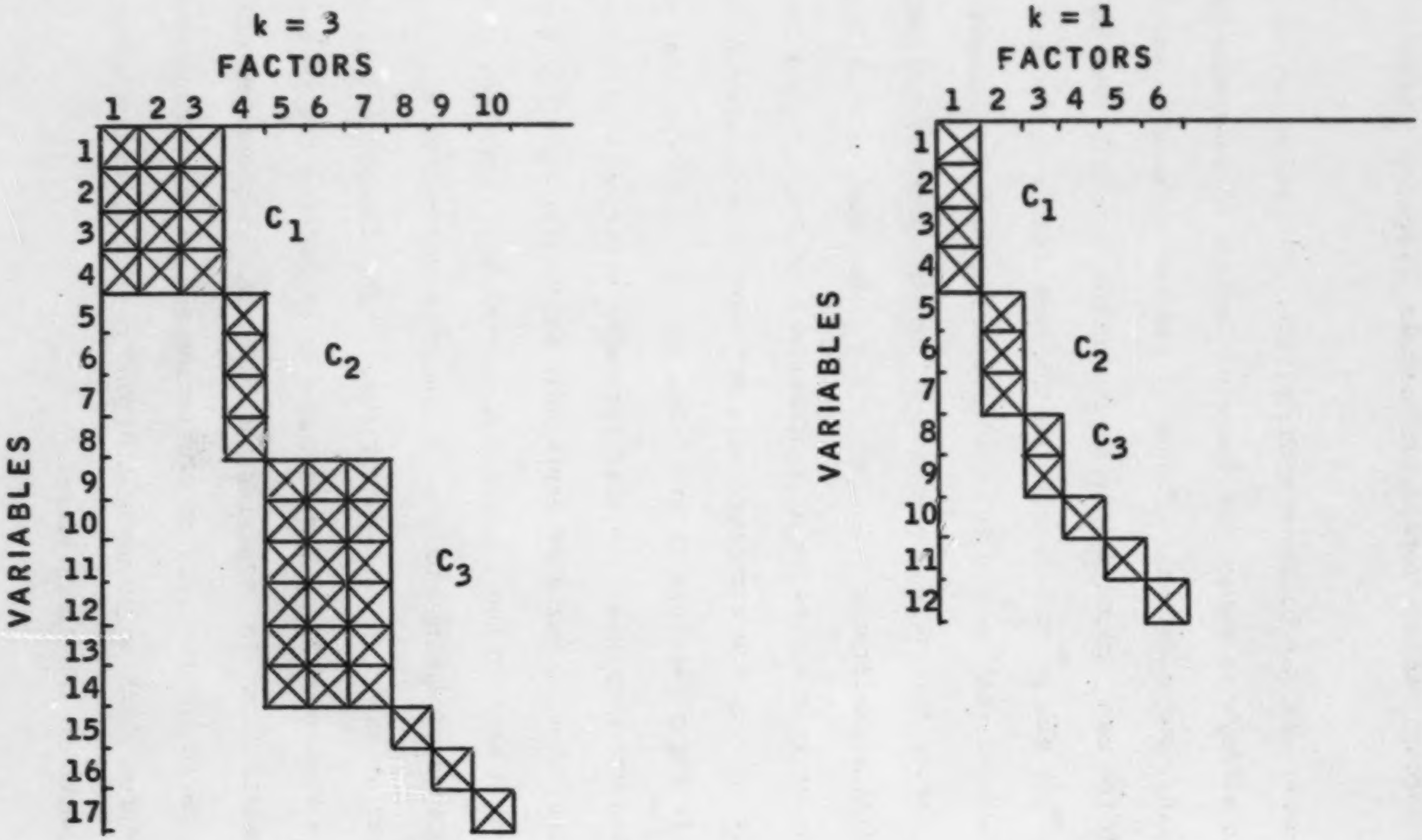

k = 3
FACTORS
1 2 3 4 5 6 7 8 9 10
VARIABLES
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
C_1
C_2
C_3
k = 1
FACTORS
1 2 3 4 5 6
VARIABLES
1 2 3 4 5 6 7 8 9 10 11 12
C_1
C_2
C_3

Fig. 4 - Disjointed Factor Spaces for k = 3 and k = 1

One can repeat the same process within each cluster and obtain subclusters. Each time the method applied would be similar to the C* method except that there is more computation involved (k-tuplets instead of pairs; $D_{(r,p,t,\ldots)}$ values instead of D_{ij} values; implicit formation of many more combinations). But this seems theoretically feasible especially through the use of sampling methods.

Let us illustrate the method in the case of $k = 2$; that is the factor space is composed of several completely disjointed 2 factor subspaces. This is a situation where triplets of variables exhaust the information available except for specific factors. This means the basic data is now composed of a three-dimensional array of values $\{D_{ijk}\}$. These values are the differences between a preassigned constant and a measure of the predictability of any variable in the triplet by the other two variables. The values D_{ijk} are positive if we group the triplet of variables in one cluster and negative otherwise. This predictability would be perfect if each variable is a linear function of only 2 factors, or in other words if they are part of a subspace generated by only 2 factors. In this case, D_{ijk} assumes its maximum positive value. This can be seen by applying an inverse transformation to any 2 variables of the triplet to generate the 2 factors and therefore the third variable. On the other hand, there exists a subspace generated by 2 factors (first 2 principal components) that would explain at least the sum of the variances, call it V, of the 2 variables in the triplet each having a variance greater than that of the third variable, whose variance we label v. Therefore the only meaningful subspace which serves as a basis is one constituted by these factors. Should the clustering index attain values which would explain $V + \epsilon$ $(\epsilon \geq 0)$ of the total variance which is $V + v$, clustering of the three variables is indicated. Otherwise, clustering is not indicated.

By computing the sum of the first two eigen-values of the covariance matrix of the three variables one obtains the variance explained by the first 2 principal components or principal factors of the 3 variables. It is known that these components are the "best" 2 factors one could obtain for the three variables. Therefore just compare the sum

$$\lambda = \lambda_1 + \lambda_2$$

where λ_1 and λ_2 are the eigen-values, to the critical quantity $V + \epsilon$ and let $D_{ijk} = \lambda - (V+\epsilon)$ if the variables considered are X_i, X_j and X_k. Thus positive D_{ijk} indicates inclusion in a cluster, negative D_{ijk} otherwise. A reasonable ϵ in this situation seems to be one where

$$\epsilon = (.7)^3 v .$$

This is so if we follow the same rationale as in the case of correlation coefficients where a correlation of .7 indicates that $(.7)^2 \cong .5$ of the variation in one variable is explained by the other variable. This statement is purely intuitive and requires much more examination.

As a last word let us mention that what was previously said about factor spaces would apply to any Euclidean space (consider profiles for instance). There a "vector" would be included in a subspace if it belongs to it or is "almost attainable" in some sense from the basis of that subspace.

REFERENCES

[1] J. J. Fortier - Contributions to Item Selection. Technical Report No. 2, Laboratory for Quantitative Research in Education, Stanford University, 1962.

[2] K. J. Holzinger and H. H. Harman - Factor Analysis. University of Chicago Press, 1941.

[3] J. A. Kahl and J. A. Davis - A Comparison of Indexes of Socio-Economic Status. American Sociological Review, Vol. XXI, No. 3, 1956.

[4] R. C. Tryon - Cluster Analysis, Ann Arbor, Michigan, Edward Bros., 1939.

ASYMPTOTIC EFFICIENCIES OF SOME TESTS OF INDEPENDENCE USED IN ITEM SELECTION

by

P. K. Bhattacharya

1. Introduction.

Suppose (X,Y) follows a bivariate normal distribution on which independent observations are made, and consider the problem of testing the hypothesis that X and Y are independent, i.e., the correlation ρ between X and Y is 0, against the alternative $\rho \neq 0$, all the other parameters being unknown. It is known that for this problem the test which rejects the null hypothesis for large values of the statistic r^2 is the uniformly most powerful invariant test (under location and scale changes) on the relevant level of significance, where r is the product moment correlation between the observed values of (X,Y) (see e.g. Lehmann [6]). However, one uses the statistic $(n-2)r^2/(1-r)^2$, which is a monotonically increasing function of r^2, because the percentage points of its distribution when $\rho = 0$ is tabulated (F distribution with 1 and n-2 d.f.), n being the sample size.

Some experiments are so designed that one or both of the variates X and Y are dichotomized. In such cases the sample product moment correlation between X and Y can no longer be computed and no optimal test for independence is known. However, some large sample tests for independence on the basis of such incomplete data are available. In the case when only one variate is dichotomized, a large sample test for independence based on biserial correlation or point biserial correlation can be constructed. For this kind of data, Mosteller [7] has suggested a procedure for estimating ρ making use of only the extreme observations on the continuous variate. With the help of this estimator, another large sample test for independence can be constructed. When both the variates are dichotomized, one can either forget all about the underlying bivariate normality and use the test for independence in a 2 x 2 contingency table or try to make use of the knowledge of bivariate normality by using the tetrachoric correlation as the test criterion.

In connection with the problem of testing for independence in a bivariate normal population on the basis of such incomplete data, the following questions naturally arise:

a) How much information is lost by dichotomizing one or both variates and using one of the tests mentioned above instead of getting the complete data and using the test based on the sample product - moment correlation. This question becomes relevant when the experimenter has a choice between taking a smaller number of observations on the original (X,Y)-variates and taking a larger number of observations on the dichotomized variates.

b) How to dichotomize the variates. This question becomes relevant when the experimenter has to dichotomize one or both variates but has a choice regarding the dividing points that will define the dichotomies.

c) When both variates are dichotomized, whether to use the test for independence in a 2 x 2 contingency table or to use the test based on tetrachoric correlation.

Since all the tests for independence based on incomplete data that are being discussed are large sample tests, we shall try to get answers to these questions in terms of asymptotic properties of these tests. Several theories have been developed for comparing large sample performances of two tests and we shall make this study in the light of two of these theories, viz. Pitman's and Bahadur's. Roughly speaking, Pitman's theory is designed to find out the efficiency of a test for detecting small departures from the null hypothesis in the large sample, whereas Bahadur's theory is designed to give an approximate idea about how good a test is for detecting arbitrary departures from the null hypothesis in the large sample.

The motivation for this study is the examination of several correlation statistics used by those working in test design for making decisions on which items should be selected for a proposed test (e.g. aptitude test, personality inventory etc.). Several of the results are of interest. It is found that the optimum dividing point for defining a dichotomy on a variate is the mean of the variate. For detecting small departures from the hypothesis of independence, an optimal dichotomy on a single variate results in about 36% loss of information, while optimal dichotomy used on both variates results in about 60% loss of information. For near-optimal dichotomy on a single variate, the test based on point biserial correlation is more efficient than the test suggested by Mosteller which is very similar to Flanagan's method. For optimal dichotomy on both variates, the 2×2 contingency χ^2 test is more efficient than the test based on tetrachoric correlation for detecting large departures from the hypothesis of independence, and for detecting small departures these two tests are equally efficient.

Several estimates for bivariate normal correlation coefficient are described in section 2. In section 3, the asymptotic distributions of these estimates have been discussed, a summary of which is given at the end of this section, and large sample tests for independence based on these estimates have been described. In section 4, the concepts of Pitman-efficiency and Bahadur-efficiency are explained briefly and formulas for these efficiencies are derived for the tests described in section 3. In section 5, tables are given for the Pitman-efficiency and Bahadur-efficiency of these tests and some general conclusions are reached on the basis of these tables.

2. Estimates for the correlation coefficient of a bivariate normal population based on incomplete data.

$(X_1,Y_1),\ldots,(X_n,Y_n),\ldots$ is a sequence of independent observations on (X,Y) which follows bivariate normal distribution with $EX = \xi$, $EY = \eta$, $\mathrm{Var}\,X = \sigma^2$, $\mathrm{Var}\,Y = \tau^2$ and $\mathrm{cov}\,(X,Y) = \rho\sigma\tau$, i.e., having a joint density

$$\frac{1}{\sigma\tau}\,\psi\left(\frac{x-\xi}{\sigma}, \frac{y-\eta}{\tau}, \rho\right)$$

where

$$\psi(x,y,\rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}(x^2-2\rho xy+y^2)\right].$$

Let μ and ν be real numbers which are the dividing points defining dichotomies in the X-variate and the Y-variate respectively. When only one variate, say the Y-variate, is dichotomized in this way, instead of the original observations (X_i,Y_i), $i=1,2,\ldots,$ we get only X_i and the knowledge whether $Y_i \leq \nu$ or $Y_i > \nu$, $i=1,2,\ldots$. When both the variates are dichotomized in the above manner, instead of the original observations (X_i,Y_i), $i=1,2,\ldots,$ we only know which of the following categories (X_i,Y_i) belongs to:

$$X_i \leq \mu \text{ and } Y_i \leq \nu ,$$
$$X_i \leq \mu \text{ and } Y_i > \nu ,$$
$$X_i > \mu \text{ and } Y_i \leq \nu ,$$
$$X_i > \mu \text{ and } Y_i > \nu .$$

For simplicity of notation it will be convenient to regard the dichotomized variates in the following way. We define

$$U_i = \begin{cases} 1 & \text{if } X_i \leq \mu \\ 0 & \text{if } X_i > \mu \end{cases} \quad \text{and} \quad V_i = \begin{cases} 1 & \text{if } Y_i \leq \nu \\ 0 & \text{if } Y_i > \nu , \end{cases}$$

$i=1,2,\ldots$. Thus, when only the Y-variate is dichotomized, the original sequence of observations $(X_1,Y_1),\ldots,(X_n,Y_n),\ldots$ is replaced by $(X_1,V_1),\ldots,(X_n,V_n),\ldots$, and when both variates are dichotomized, the original sequence of observations is replaced by $(U_1,V_1),\ldots,(U_n,V_n),\ldots$.

On the basis of the first n of the original sequence of observations, ρ is estimated by the statistic

$$r_n^{(1)} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_iY_i - \overline{X}_n\overline{Y}_n}{[(\frac{1}{n}\sum_{i=1}^{n} X_i^2-\overline{X}_n^2)\cdot(\frac{1}{n}\sum_{i=1}^{n} Y_i^2-\overline{Y}_n^2)]^{1/2}} ,$$

where $\overline{X}_n = \sum_{i=1}^{n} X_i/n$ and $\overline{Y}_n = \sum_{i=1}^{n} Y_i/n$. We know that $r_n^{(1)}$ is the maximum likelihood estimate of ρ based on $(X_1,Y_1),\ldots,(X_n,Y_n)$ and is consistent.

Let us define $\alpha = (\mu-\xi)/\sigma$, $\beta = (\nu-\eta)/\tau$; α and β are the standardized values of the dividing points μ and ν respectively, which define the dichotomies of the original variates. We shall denote by $\varphi(x)$ the standard normal density function and by $\Phi(x)$ the standard normal cumulative distribution function, i.e.

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

and

$$\Phi(x) = \int_{-\infty}^{x} \varphi(t)dt \ .$$

Now consider the problem of estimating ρ on the basis of the first n of the sequence of observations in which the Y-variate has been dichotomized. For the time being let us suppose that σ and β are known. If $(X_1,Y_1),\ldots,(X_n,Y_n)$ were available, a very crude method for estimating the slope of the regression line of X given Y would be to compute the mean of all Y-observations less than ν and the mean of all X-observations associated with these Y-observations and to plot these means $(\overline{Y}_n^{(1)}, \overline{X}_n^{(1)})$ on the (Y,X)-plane. Similarly, the mean $\overline{Y}_n^{(2)}$ of all Y-observations greater than ν and the mean $\overline{X}_n^{(2)}$ of all X-observations associated with these Y-observations, are also represented as a point $(\overline{Y}_n^{(2)}, \overline{X}_n^{(2)})$ on the (Y,X)-plane. The slope of the line joining these two points, would be

$$\frac{\overline{X}_n^{(2)} - \overline{X}_n^{(1)}}{\overline{Y}_n^{(2)} - \overline{Y}_n^{(1)}} ,$$

and would be a rough estimate of the slope of the regression line of X given Y, i.e., of $\rho\sigma/\tau$. However, the denominator of this estimate cannot be computed from the data in which the Y-variate is dichotomized. But for large n, $\overline{Y}_n^{(1)}$ would be approximately equal to the conditional expectation of Y given $Y \leq \nu$ and $\overline{Y}_n^{(2)}$ would be approximately equal to the conditional expectation of Y given $Y > \nu$. These conditional expectations are $\eta - \tau \cdot \frac{\varphi(\beta)}{\Phi(\beta)}$ and $\eta + \tau \cdot \frac{\varphi(\beta)}{1-\Phi(\beta)}$ respectively. Hence

$$\overline{Y}_n^{(2)} - \overline{Y}_n^{(1)} \simeq \tau \cdot \frac{\varphi(\beta)}{1-\Phi(\beta)} + \tau \cdot \frac{\varphi(\beta)}{\Phi(\beta)} = \frac{\tau\,\varphi(\beta)}{\Phi(\beta)\{1-\Phi(\beta)\}} .$$

Thus we have $\frac{\overline{X}_n^{(2)}-\overline{X}_n^{(1)}}{\tau} \cdot \frac{\Phi(\beta)\{1-\Phi(\beta)\}}{\varphi(\beta)}$ as an estimate of $\rho\sigma/\tau$, and so

$$\frac{\overline{X}_n^{(2)}-\overline{X}_n^{(1)}}{\sigma} \cdot \frac{\Phi(\beta)\{1-\Phi(\beta)\}}{\varphi(\beta)}$$

is a rough estimate of ρ. So far we have assumed that σ and β are known. Now let us remove this assumption and estimate them from $(X_1,V_1),\ldots,(X_n,V_n)$. Since σ is the standard deviation of X in the population, it is estimated by the sample standard deviation of X, viz. $[\frac{1}{n}\sum_{i=1}^{n} X_i^2-\overline{X}_n^2]^{1/2}$. Also, $\Phi(\beta)$ is the probability that $Y \leq \nu$, and is estimated by the proportion of $Y \leq \nu$ in the sample. Since V_i attains the value 1 when $Y_i \leq \nu$ and the value 0 when $Y_i > \nu$, this sample proportion is expressed as $\overline{V}_n$ (we are denoting $\overline{U}_n = \sum_{i=1}^{n} U_i/n$ and $\overline{V}_n = \sum_{i=1}^{n} V_i/n$). Thus, ρ is estimated by

$$r_n^{(2)} = \frac{\overline{X}_n^{(2)} - \overline{X}_n^{(1)}}{[\frac{1}{n} \sum_{i=1}^{n} X_i^2 - \overline{X}_n^2]^{1/2}} \cdot \frac{\overline{V}_n(1-\overline{V}_n)}{\varphi(\Phi^{-1}(\overline{V}_n))} ,$$

where $\Phi^{-1}(p)$ is that value of x for which $\Phi(x) = p$, $0 < p < 1$. Now $\overline{X}_n^{(1)}$ and $\overline{X}_n^{(2)}$ can be expressed in terms of $(X_1,V_1),\ldots,(X_n,V_n)$. $\overline{X}_n^{(2)}$ is the mean of the X-observations associated with Y-observations less than ν, i.e., with V-observations equal to 1. The sum of these X-observations is $\sum_{i=1}^{n} X_i V_i$ and the total number of such observations is $\sum_{i=1}^{n} V_i$. Hence

$$\overline{X}_n^{(1)} = \sum_{i=1}^{n} X_i V_i \Big/ \sum_{i=1}^{n} V_i .$$

Similarly,

$$\overline{X}_n^{(2)} = \sum_{i=1}^{n} X_i (1-V_i)/(n - \sum_{i=1}^{n} V_i) ,$$

and thus we have

$$\overline{X}_n^{(2)} - \overline{X}_n^{(1)} = - \frac{\frac{1}{n} \sum_{i=1}^{n} X_i V_i - \overline{X}_n \overline{V}_n}{\overline{V}_n(1-\overline{V}_n)} ,$$

and finally, the estimate for ρ is given by

$$r_n^{(2)} = - \frac{\frac{1}{n} \sum_{i=1}^{n} X_i V_i - \overline{X}_n \overline{V}_n}{[\frac{1}{n} \sum_{i=1}^{n} X_i^2 - \overline{X}_n^2]^{1/2}} \cdot \frac{1}{\varphi(\Phi^{-1}(\overline{V}_n))} .$$

Now $\overline{V}_n(1-\overline{V}_n) = \overline{V}_n - \overline{V}_n^2 = \frac{1}{n} \sum_{i=1}^{n} V_i - \overline{V}_n^2 = \frac{1}{n} \sum_{i=1}^{n} V_i^2 - \overline{V}_n^2$ since $V_i = V_i^2$, the only possible values of V_i being 0 and 1. Hence this estimate can be written as

$$r_n^{(2)} = - \frac{\frac{1}{n}\sum_{i=1}^{n} X_i V_i - \bar{X}_n \bar{V}_n}{[(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2)\cdot(\frac{1}{n}\sum_{i=1}^{n} V_i^2 - \bar{V}_n^2)]^{1/2}} \cdot \frac{[\bar{V}_n(1-\bar{V}_n)]^{1/2}}{\varphi(\Phi^{-1}(\bar{V}_n))} .$$

This statistic is known as biserial correlation and was introduced by K. Pearson [10]. This statistic can have values beyond $[-1, +1]$ in the small sample but as we shall see in the next section, it is a consistent estimate of ρ.

If we compute the correlation coefficient between X and V based on $(X_1,V_1),\ldots,(X_n,V_n)$, then that statistic

$$r_n^{(3)} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i V_i - \bar{X}_n \bar{V}_n}{[(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2)\cdot(\frac{1}{n}\sum_{i=1}^{n} V_i^2 - \bar{V}_n^2)]^{1/2}}$$

is known as point biserial correlation. Comparing $r_n^{(3)}$ with $r_n^{(2)}$ we see that

$$r_n^{(3)} = - r_n^{(2)} \cdot \frac{\varphi(\Phi^{-1}(\bar{V}_n))}{[\bar{V}_n(1-\bar{V}_n)]^{1/2}} .$$

We have already mentioned that $r_n^{(2)}$ is a consistent estimate of ρ and $\bar{V}_n$ is a consistent estimate of $\Phi(\beta)$. Hence $r_n^{(3)}$ is not a consistent estimate of ρ but a consistent estimate of

$$- \frac{\rho\, \varphi(\beta)}{[\Phi(\beta)\{1-\Phi(\beta)\}]^{1/2}} .$$

Thus $r_n^{(3)}$ estimates ρ only indirectly in the sense that it estimates the function of ρ given above. But this function vanishes only when $\rho = 0$ and for this reason it can be used for constructing a large sample test for $H_o:\rho = 0$. Point biserial correlation usually occurs in statistical literature in connection with a different model. Suppose $(X_1,V_1),\ldots,(X_n,V_n)$ are independent observations from a population with $P[V=1] = \pi$, $P[V=0] = 1-\pi$, and the conditional distribution of X given $V=i$ is normal with mean ξ_i and variance σ^2, $i=0,1$. Then the point biserial correlation $r_n^{(3)}$ between the X and V observations is used for testing the hypothesis that $\xi_1 = \xi_2$ (see e.g. Tate [11]). However, we are going to use $r_n^{(3)}$ in connection with an altogether different model and in the next section we shall study its limiting distribution under the appropriate model.

We shall now consider the special case when ξ, η, σ^2 and τ are known and ν (the dividing point which defines the dichotomy in the Y-variate) is equal to η. For simplicity we assume (without any loss of generality) that $\xi=\eta=0$, $\sigma^2=\tau^2=1$ and $\nu=0$. Let us consider the following regions in the (X,Y)-plane:

Region 1. $X \geq k$, $Y \geq 0$
Region 2. $X \leq -k$, $Y \geq 0$
Region 3. $X \leq -k$, $Y \leq 0$
Region 4. $X \geq k$, $Y < 0$
Region 5. $-k < X < k$

These regions are non-overlapping and exhaustive. Let us denote by $P_i(\rho)$, the probability of a sample point being included in the region i. Then

$$P_1(\rho) = \int_k^\infty \int_0^\infty \psi(x,y,\rho)dxdy \ , \quad P_2(\rho) = \int_{-\infty}^{-k} \int_0^\infty \psi(x,y,\rho)dxdy \ ,$$

$$P_3(\rho) = \int_{-\infty}^{-k} \int_{-\infty}^{0} \psi(x,y,\rho)dxdy \ , \quad P_4(\rho) = \int_k^\infty \int_{-\infty}^{0} \psi(x,y,\rho)dxdy \ ,$$

and $$P_5(\rho) = \int_{-k}^{k} \varphi(x)dx \ .$$

This last probability $P_5(\rho)$ is the same for all ρ. Now the data $(X_1,V_1),\ldots,(X_n,V_n)$ permits us to find the number of original observations $(X_1,Y_1),\ldots,(X_n,Y_n)$ belonging to each of the five regions on the (X-Y)-plane mentioned above. Let n_i be the number of points belonging to Region i , $\sum_{i=1}^{5} n_i = n$. When ρ obtains, the joint probability distribution of the n_i's is

$$\frac{n!}{\prod_{i=1}^{5} n_i!} \prod_{i=1}^{5} [P_i(\rho)]^{n_i} \ .$$

This likelihood function attains its maximum at $\hat{\rho}_n$ where $\hat{\rho}_n$ is the solution of the equation

$$P_1(\rho) = (1-\Phi(k)) \cdot \frac{n_1+n_3}{n_1+n_2+n_3+n_4}$$

in ρ. Nothing has been said so far about k which is at our choice. Now the variance of $\hat{\rho}_n$ when $\rho = 0$, attains its minimum when $k = .6121$ and $1-\Phi(k) = .2702$ ($-k$ and k are approximately the 27th and 73rd percentile points of the marginal distribution of X). Let us denote by $r_n^{(4)}$, the estimate of ρ obtained with the optimal choice of $k = .6121$. This method of estimating ρ has been suggested by Mosteller [7]. Similar methods have also been suggested by Kelley [4] and Flanagan [3].

We shall now consider the problem of estimating ρ when both variates are dichotomized, i.e., we shall try to estimate ρ from $(U_1,V_1),\ldots,(U_n,V_n)$. Let

$$I(\alpha,\beta,\rho) = \int_{-\infty}^{\alpha}\int_{-\infty}^{\beta} \psi(x,y,\rho)dxdy ,$$

the probability that $X_i \leq \mu$ and $Y_i \leq \nu$. If α, β and $I(\alpha,\beta,\rho)$ were known, then ρ could be obtained. But it has already been mentioned that $\Phi(\alpha)$, the probability of $X \leq \mu$ is estimated by $\overline{U}_n$, the sample proportion of $X_i \leq \mu$, $\Phi(\beta)$, the probability of $Y \leq \nu$ is estimated by $\overline{V}_n$, the sample proportion of $Y_i \leq \nu$, and $I(\alpha,\beta,\rho)$, the probability of the joint event $X \leq \mu$ and $Y \leq \nu$ is estimated by the proportion of cases in the sample for which $X_i \leq \mu$ and $Y_i \leq \nu$. To express this last sample proportion in terms of $(U_1,V_1),\ldots,(U_n,V_n)$, we note that the event $X_i \leq \mu$ and $Y_i \leq \nu$ is equivalent to the event $U_i = 1$ and $V_i = 1$ which in its turn is equivalent to the event $W_i = U_iV_i = 1$. Hence $I(\alpha,\beta,\rho)$ is estimated by $\overline{W}_n = \sum_{i=1}^{n} W_i/n$. Again since $\Phi(\alpha)$ and $\Phi(\beta)$ are estimated by $\overline{U}_n$ and $\overline{V}_n$ respectively, α and β are estimated by

$$a_n = \Phi^{-1}(\overline{U}_n) \text{ and } b_n = \Phi^{-1}(\overline{V}_n)$$

respectively. We can now solve the equation

$$I(a_n,b_n,r) = \overline{W}_n$$

in r to get an estimate $r_n^{(5)}$ of ρ. This estimate is known as tetrachoric correlation and was introduced by K. Pearson [9]. $r_n^{(5)}$ can be obtained from a_n, b_n and $\overline{W}_n$ by inverse interpolation using the tables for I(a,b,r). (For other methods of computing $r_n^{(5)}$, see Kendall [5].) As will be seen in the next section, $r_n^{(5)}$ is a consistent estimate of ρ.

Now let us consider the correlation coefficient between U and V based on $(U_1,V_1),\ldots,(U_n,V_n)$, i.e., the statistic

$$r_n^{(6)} = \frac{\frac{1}{n}\sum_{i=1}^{n} U_iV_i - \overline{U}_n\overline{V}_n}{[(\frac{1}{n}\sum_{i=1}^{n} U_i^2-\overline{U}_n^2)\cdot(\frac{1}{n}\sum_{i=1}^{n} V_i^2-\overline{V}_n^2)]^{1/2}} .$$

$r_n^{(6)}$ is also known in statistical literature as Phi coefficient.

Define $n_{00} = \sum_{i=1}^{n} U_iV_i$, $n_{01} = \sum_{i=1}^{n} U_i - n_{00}$, $n_{10} = \sum_{i=1}^{n} V_i - n_{00}$ and $n_{11} = n-n_{00}-n_{01}-n_{10}$. Then after some simplification we get

$$nr_n^{(6)^2} = \frac{(n_{00}\,n_{11} - n_{10}\,n_{01})^2}{(n_{00}+n_{01})(n_{00}+n_{10})(n_{01}+n_{11})(n_{10}+n_{11})}$$

$$= \sum_{j=0}^{1}\sum_{k=0}^{1} \frac{[n_{jk}-n\,(\frac{n_{j0}+n_{j1}}{n})\cdot(\frac{n_{0k}+n_{1k}}{n})]^2}{n(\frac{n_{j0}+n_{j1}}{n})\cdot(\frac{n_{0k}+n_{1k}}{n})} .$$

The last two expressions are well-known formulas for the statistic used for testing independence in a 2 x 2 contingency table. Like $r_n^{(3)}$, $r_n^{(6)}$ is not a consistent estimate of ρ but it is a consistent estimate of

$$\frac{I(\alpha,\beta,\rho) - \Phi(\alpha)\Phi(\beta)}{[\Phi(\alpha)\{1-\Phi(\alpha)\}\Phi(\beta)\{1-\Phi(\beta)\}]^{1/2}} .$$

Thus $r_n^{(6)}$ estimates ρ in an indirect manner in the sense that it estimates the above function of ρ. But this function vanishes only when $\rho=0$ and therefore the statistic $r_n^{(6)}$ or $nr_n^{(6)2}$ can be used to construct a large sample test for $H_o:\rho=0$.

3. Asymptotic distribution of the statistics mentioned in section 2 and large sample tests of $H_o:\rho=0$ based on these statistics.

We first note that each of the statistics $r_n^{(1)}$, $r_n^{(3)}$ and $r_n^{(6)}$ is a correlation coefficient between two sets of observations. For this reason their asymptotic distributions can be studied in the same way.

Let us first consider $r_n^{(1)}$ and suppose $\rho=0$.

$$\sqrt{n}\; r_n^{(1)} = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_iY_i - \sqrt{n}\, \bar{X}_n\bar{Y}_n}{[(\frac{1}{n} \sum_{i=1}^{n} X_i^2-\bar{X}_n^2)\cdot(\frac{1}{n} \sum_{i=1}^{n} Y_i^2-\bar{Y}_n^2)]^{1/2}} .$$

Now

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_iY_i - \sqrt{n}\; \bar{X}_n\bar{Y}_n$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{(X_i-\xi)+\xi\}\{(Y_i-\eta)+\eta\} - \sqrt{n}\; \{(\bar{X}_n-\xi)+\xi\}\{(\bar{Y}_n-\eta)+\eta\}$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i-\xi)(Y_i-\eta) - \frac{1}{\sqrt{n}}\; \{\sqrt{n}\; (\bar{X}_n-\xi)\}\cdot\{\sqrt{n}\; (\bar{Y}_n-\eta)\} .$$

It follows from the central limit theorem that

(i) The distribution of $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i-\xi)(Y_i-\eta)$ approaches normal distribution with mean 0 and variance $\sigma^2\tau^2$ since $\rho=0$.

(ii) $\sqrt{n}\ (\overline{X}_n-\xi)$ is asymptotically normal with mean 0 and variance σ^2 and $\sqrt{n}\ (\overline{Y}_n-\eta)$ is asymptotically normal with mean 0 and variance τ^2. Hence $\frac{1}{\sqrt{n}}\ \{\sqrt{n}\ (\overline{X}_n-\xi)\}\cdot\{\sqrt{n}\ (\overline{Y}_n-\eta)\}$ approaches zero in probability.

(iii) $\frac{1}{n} \sum_{i=1}^{n} X_i^2-\overline{X}_n^2$ and $\frac{1}{n}\ Y_i^2-\overline{Y}_n^2$ are consistent estimates of σ^2 and τ^2 respectively.

From (i), (ii) and (iii) it now follows that when $\rho=0$, $\sqrt{n}\ r_n^{(1)}$ is asymptotically normal with mean 0 and variance 1. By an application of a theorem of Cramér [2] we finally get the asymptotic distribution of $\sqrt{n}\ r_n^{(1)} \Big/ \sqrt{1-r_n^{(1)^2}}$ under the null hypothesis as normal with mean 0 and variance

$$\left[\frac{d}{d\rho}\left(\frac{\rho}{\sqrt{1-\rho^2}}\right)\right]_{\rho=0}$$

which is 1. Thus the statistic

$$T_n^{(1)} = \frac{n\, r_n^{(1)^2}}{1-r_n^{(1)^2}}$$

has limiting χ^2 distribution with 1 d.f. when $\rho=0$.

To find the asymptotic distribution of $r_n^{(1)}$ when $\rho \neq 0$, we first note that

$$(\sqrt{n}\ [(\frac{1}{n}\sum_{i=1}^{n} X_i^2-\overline{X}_i^2)-\sigma^2]\ ,\ \sqrt{n}\ [(\frac{1}{n}\sum_{i=1}^{n} Y_i^2-\overline{Y}_n^2)-\tau^2]\ ,$$

$$\sqrt{n}\ [(\frac{1}{n}\sum_{i=1}^{n} X_iY_i-\overline{X}_n\overline{Y}_n)-\rho\sigma\tau])$$

has asymptotic multivariate normal distribution with mean $(0,0,0)$ and some covariance matrix which we do not have to worry about. By an application of Cramér's theorem it follows that

$$\sqrt{n}\left(\frac{r_n^{(1)}}{\sqrt{1-r_n^{(1)^2}}} - \frac{\rho}{\sqrt{1-\rho^2}}\right)$$

has normal distribution with mean zero and some variance which we do not need.

Consider next the statistic

$$\sqrt{n}\ r_n^{(3)} = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_iV_i - \overline{X}_n\overline{V}_n}{[(\frac{1}{n}\sum_{i=1}^{n} X_i^2-\overline{X}_n^2)\cdot(\frac{1}{n}\sum_{i=1}^{n} V_i^2-\overline{V}_n^2)]^{1/2}}\ .$$

Again we note that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_iV_i - \overline{X}_n\overline{V}_n$$

$$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} (X_i-\xi)(V_i-\Phi(\beta)) - \frac{1}{\sqrt{n}}\ \{\sqrt{n}\ (\overline{X}_n-\xi)\}\cdot\{\ \sqrt{n}\ (\overline{V}_n-\Phi(\beta))\}\ .$$

When $\rho=0$,

(i) $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} (X_i-\xi)(V_i-\Phi(\beta))$ has asymptotic normal distribution with mean 0 and variance $\sigma^2\Phi(\beta)(1-\Phi(\beta))$.

(ii) $\sqrt{n}\,(\overline{X}_n-\xi)$ is asymptotically normal with mean 0 and variance σ^2 and $\sqrt{n}\,(\overline{V}_n-\Phi(\beta))$ is asymptotically normal with mean 0 and variance $\Phi(\beta)(1-\Phi(\beta))$. Hence $\frac{1}{\sqrt{n}}\,\{\sqrt{n}\,(\overline{X}_n-\xi)\}\cdot\{\sqrt{n}\,(\overline{V}_n-\Phi(\beta))\}$ approaches zero in probability.

(iii) $\frac{1}{n}\sum_{i=1}^{n} X_i^2-\overline{X}_n^2$ and $\frac{1}{n}\sum_{i=1}^{n} V_i^2-\overline{V}_n^2$ are consistent estimates of σ^2 and $\Phi(\beta)(1-\Phi(\beta))$ respectively.

From (i), (ii) and (iii) it now follows that when $\rho=0$, $\sqrt{n}\,r_n^{(3)}$ is asymptotically normal with mean 0 and variance 1. Thus the statistic $T_n^{(3)} = nr_n^{(3)^2}$ has a limiting χ^2 distribution with 1 d.f. under the null hypothesis.

To find the asymptotic distribution of $r_n^{(3)}$ when $\rho \neq 0$, we note that

$$\Big(\sqrt{n}\,[(\frac{1}{n}\sum_{i=1}^{n} X_i^2-\overline{X}_n^2)-\sigma^2]\,,\ \sqrt{n}\,[(\frac{1}{n}\sum_{i=1}^{n} V_i^2-\overline{V}_n^2)-\Phi(\beta)(1-\Phi(\beta))]\,,$$

$$\sqrt{n}\,[(\frac{1}{n}\sum_{i=1}^{n} X_iV_i-\overline{X}_n\overline{V}_n)+\rho\sigma\varphi(\beta)]\Big)$$

has asymptotic multivariate normal distribution with mean (0,0,0) and some covariance matrix. This is true because

$$E[(X-\xi)(V-\Phi(\beta))]$$

$$= E(XV)-\xi\Phi(\beta)$$

$$= \int_{-\infty}^{\infty} x \cdot \frac{1}{\sigma} \varphi(\frac{x-\xi}{\sigma})\Phi(\frac{\beta}{(1-\rho^2)^{1/2}} - \frac{\rho}{(1-\rho^2)^{1/2}} (\frac{x-\xi}{\sigma}))dx-\xi\Phi(\beta)$$

$$= \int_{-\infty}^{\infty} (\xi+\sigma x)\varphi(x)\Phi(\frac{\beta}{(1-\rho^2)^{1/2}} - \frac{\rho\, x}{(1-\rho^2)^{1/2}})dx -\xi\Phi(\beta)$$

$$= \xi\int_{-\infty}^{\infty} \varphi(x)\Phi(\frac{\beta}{(1-\rho^2)^{1/2}} - \frac{\rho\, x}{(1-\rho^2)^{1/2}})dx +$$

$$\sigma\int_{-\infty}^{\infty} x\, \varphi(x)\Phi(\frac{\beta}{(1-\rho^2)^{1/2}} - \frac{\rho\, x}{(1-\rho^2)^{1/2}})dx-\xi\Phi(\beta)$$

$$= \xi\Phi(\beta)-\rho\sigma\varphi(\beta)-\xi\Phi(\beta)$$

$$= - \rho\sigma\varphi(\beta) .$$

It now follows from Cramér's theorem that $\sqrt{n}\ (r_n^{(3)} + \frac{\rho\ \varphi(\beta)}{[\Phi(\beta)(1-\Phi(\beta))]^{1/2}})$ has asymptotic normal distribution with mean 0 and some variance that we do not need.

To find the asymptotic distribution of

$$\sqrt{n}\ r_n^{(6)} = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_iV_i - \overline{U}_n\overline{V}_n}{[(\frac{1}{n} \sum_{i=1}^{n} U_i^2-\overline{U}_n^2)\cdot(\frac{1}{n} \sum_{i=1}^{n} V_i^2-\overline{V}_n^2)]^{1/2}}$$

when $\rho=0$ we first note that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_iV_i - \overline{U}_n\overline{V}_n$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (U_i-\Phi(\alpha))(V_i\Phi(\beta)) - \frac{1}{\sqrt{n}} \{\sqrt{n}\ (\overline{U}_n-\Phi(\alpha))\}\cdot\{\sqrt{n}\ (\overline{V}_n-\Phi(\beta))\} .$$

We now see that

(i) $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (U_i-\Phi(\alpha))(V_i-\Phi(\beta))$ has asymptotic normal distribution with mean 0 and variance $\Phi(\alpha)(1-\Phi(\alpha))\Phi(\beta)(1-\Phi(\beta))$.

(ii) $\sqrt{n}\ (\overline{U}_n-\Phi(\alpha))$ has asymptotic normal distribution with mean 0 and variance $\Phi(\alpha)(1-\Phi(\alpha))$ and $\sqrt{n}\ (\overline{V}_n-\Phi(\beta))$ has asymptotic normal distribution with mean 0 and variance $\Phi(\beta)(1-\Phi(\beta))$. Hence $\frac{1}{\sqrt{n}} \{\sqrt{n}\ (\overline{U}_n-\Phi(\alpha))\}\cdot\{\sqrt{n}\ (\overline{V}_n-\Phi(\beta))\}$ approaches zero in probability .

(iii) $\frac{1}{n} \sum_{i=1}^{n} U_i^2-\overline{U}_n^2$ and $\frac{1}{n} \sum_{i=1}^{n} V_i^2-\overline{V}_n^2$ are consistent estimates of $\Phi(\alpha)(1-\Phi(\alpha))$ and $\Phi(\beta)(1-\Phi(\beta))$ respectively.

From (i), (ii), and (iii) it follows that when $\rho=0$, $\sqrt{n}\ r_n^{(6)}$ is asymptotically normal with mean 0 and variance 1. Thus the statistic $T_n^{(6)} = nr_n^{(6)2}$ has a limiting χ^2 with 1 d.f. under the null hypothesis.

To find the asymptotic distribution of $r_n^{(6)}$ when $\rho \neq 0$, we first note that

$$(\sqrt{n}\ [(\frac{1}{n} \sum_{i=1}^{n} U_i^2-\overline{U}_n^2)-\Phi(\alpha)(1-\Phi(\alpha))]\ ,\ \sqrt{n}\ [(\frac{1}{n} \sum_{i=1}^{n} V_i^2-\overline{V}_n^2)-\Phi(\beta)(1-\Phi(\beta))]\ ,$$

$$\sqrt{n}\ [(\frac{1}{n} \sum_{i=1}^{n} U_iV_i-\overline{U}_n\overline{V}_n)-(I(\alpha,\beta,\rho)-\Phi(\alpha)\Phi(\beta))])$$

has asymptotic multivariate normal distribution with mean (0,0,0) and some covariance matrix. This is true because,

$$\begin{aligned} & E[(U-\Phi(\alpha))(V-\Phi(\beta))] \\ &= E(UV)-\Phi(\alpha)\Phi(\beta) \\ &= P[U=1,V=1]-\Phi(\alpha)\ \Phi(\beta) \\ &= P[X \leq \mu\ ,\ Y \leq \nu]-\Phi(\alpha)\ \Phi(\beta) \\ &= I(\alpha,\beta,\rho)-\Phi(\alpha)\ \Phi(\beta)\ . \end{aligned}$$

It now follows from Cramér's theorem that

$$\sqrt{n}\ (r_n^{(6)} - \frac{I(\alpha,\beta,\rho)-\Phi(\alpha)\ \Phi(\beta)}{[\Phi(\alpha)(1-\Phi(\alpha))\Phi(\beta)(1-\Phi(\beta))]^{1/2}})$$

has asymptotic normal distribution with mean 0 and some variance which we do not need.

We have already seen in section 2 that

$$r_n^{(2)} = - r_n^{(3)} \cdot \frac{[\overline{V}_n(1-\overline{V}_n)]^{1/2}}{\varphi(\Phi^{-1}(\overline{V}_n))} \quad .$$

Since $\sqrt{n}\ r_n^{(3)}$ is asymptotically normal with mean 0 and variance 1 and since $[\overline{V}_n(1-\overline{V}_n)]^{1/2}/\varphi(\Phi^{-1}(\overline{V}_n))$ approaches $[\Phi(\beta)(1-\Phi(\beta))]^{1/2}/\varphi(\beta)$ in probability ($\overline{V}_n$ being a consistent estimate of $\Phi(\beta)$), it follows that $\sqrt{n}\ r_n^{(2)}$ is asymptotically normal with mean 0 and variance $\Phi(\beta)(1-\Phi(\beta))/\varphi^2(\beta)$. Hence

$$\frac{n r_n^{(2)^2}\ \varphi^2(\beta)}{\Phi(\beta)(1-\Phi(\beta))}$$

has χ^2 distribution with 1 d.f. under the null hypothesis. But this involves β which is unknown. If we substitute $\Phi(\beta)$ by its consistent estimate $\overline{V}_n$, we get

$$T_n^{(2)} = \frac{n\ r_n^{(2)^2}\ \varphi^2\ (\Phi^{-1}(\overline{V}_n))}{\overline{V}_n(1-\overline{V}_n)} = n\ r_n^{(3)^2} = T_n^{(3)}$$

Thus the test for independence based on $r_n^{(2)}$ turns out to be the same as the test for independence based on $r_n^{(3)}$.

When $\rho = 0$, $\sqrt{n}\, r_n^{(4)}$ is asymptotically normal with mean 0 and variance 1.939 (approximately). In other words, $\sqrt{n}\, r_n^{(4)}/\sqrt{1.939}$ is asymptotically normal with mean 0 and variance 1. Hence the statistic $T_n^{(4)} = n{r_n^{(4)}}^2/1.939$ has a limiting χ^2 distribution with 1 d.f. under the null hypothesis.

When $\rho \neq 0$, $\sqrt{n}\,(r_n^{(4)}-\rho)$ has asymptotic normal distribution with mean 0 and a variance which we do not need.

It now remains to find the limiting distribution of $r_n^{(5)}$. We first note that $\sqrt{n}\,(\overline{U}_n-\Phi(\alpha)\ ,\ \overline{V}_n-\Phi(\beta)\ ,\ \overline{W}_n-I(\alpha,\beta,\rho))$ has the multivariate normal limit distribution with mean (0,0,0) and covariance matrix

$$\begin{bmatrix} \Phi(\alpha)(1-\Phi(\alpha)) & I(\alpha,\beta,\rho)-\Phi(\alpha)\Phi(\beta) & I(\alpha,\beta,\rho)(1-\Phi(\alpha)) \\ I(\alpha,\beta,\rho)-\Phi(\alpha)\Phi(\beta) & \Phi(\beta)(1-\Phi(\beta)) & I(\alpha,\beta,\rho)(1-\Phi(\beta)) \\ I(\alpha,\beta,\rho)(1-\Phi(\alpha)) & I(\alpha,\beta,\rho)(1-\Phi(\beta)) & I(\alpha,\beta,\rho)(1-I(\alpha,\beta,\rho)) \end{bmatrix} .$$

Since $a_n = \Phi^{-1}(\overline{U}_n)$, $b_n = \Phi^{-1}(\overline{V}_n)$ and $\overline{W}_n$ have continuous first partial derivatives with respect to $\overline{U}_n$, $\overline{V}_n$ and $\overline{W}_n$ at $\overline{U}_n=\Phi(\alpha)$, $\overline{V}_n = \Phi(\beta)$ and $\overline{W}_n = I(\alpha,\beta,\rho)$, it follows from Cramér's theorem that

$$\sqrt{n}\,(a_n-\alpha,\ b_n-\beta,\ \overline{W}_n - I(\alpha,\beta,\rho))$$

has the multivariate normal limit distribution with mean (0,0,0) and covariance matrix

$$\begin{bmatrix} \sigma_{aa} & \sigma_{ab} & \sigma_{aW} \\ \sigma_{ab} & \sigma_{bb} & \sigma_{bW} \\ \sigma_{aW} & \sigma_{bW} & \sigma_{WW} \end{bmatrix},$$

where

$$\sigma_{aa} = \Phi(\alpha)(1-\Phi(\alpha))/\varphi^2(\alpha) \quad , \quad \sigma_{bb} = \Phi(\beta)(1-\Phi(\beta))/\varphi^2(\beta) \quad ,$$

$$\sigma_{WW} = I(\alpha,\beta,\rho)(1-I(\alpha,\beta,\rho)) \quad , \quad \sigma_{ab} = (I(\alpha,\beta,\rho)-\Phi(\alpha)\Phi(\beta))/\varphi(\alpha)\varphi(\beta),$$

$$\sigma_{aW} = I(\alpha,\beta,\rho)(1-\Phi(\alpha))/\varphi(\alpha) \quad , \quad \sigma_{bW} = I(\alpha,\beta,\rho)(1-\Phi(\beta))/\varphi(\beta) \; .$$

We now study the behavior of $r_n^{(5)}$ defined by

$$I(a_n, b_n, r_n^{(5)}) = \overline{W}_n \quad ,$$

as a function of a_n, b_n and $\overline{W}_n$. To indicate this dependence, we write $r_n^{(5)}$ defined by the above equation, as $r(a_n,b_n,\overline{W}_n)$. Obviously, $r(\alpha,\beta, I(\alpha,\beta,\rho)) = \rho$. For simplicity, we shall sometimes denote $I(\alpha,\beta,\rho)$ by simply I. Now,

$$\left.\frac{\partial r(a,b,W)}{\partial a}\right|_{a=\alpha,\ b=\beta,\ W=I(\alpha,\beta,\rho)}$$

$$= \lim_{\epsilon\to 0} \frac{r(\alpha+\epsilon,\beta,I(\alpha,\beta,\rho)) - r(\alpha,\beta,I(\alpha,\beta,\rho))}{\epsilon} = \lim_{\epsilon\to 0} \frac{\Delta(\epsilon)}{\epsilon} \quad \text{say .}$$

Since r is a continuous function $\lim_{\epsilon\to 0} \Delta(\epsilon) = 0$. By definition,

$$I = \int_{-\infty}^{\alpha}\int_{-\infty}^{\beta} \psi(x,y,r(\alpha,\beta,I))\, dx\, dy$$

$$= \int_{-\infty}^{\alpha+\epsilon}\int_{-\infty}^{\beta} \psi(x,y,r(\alpha+\epsilon,\beta,I))\, dxdy$$

$$= \int_{-\infty}^{\alpha+\epsilon}\int_{-\infty}^{\beta} [\psi(x,y,r(\alpha,\beta,I)) + \Delta(\epsilon)\left\{\frac{\partial\psi(x,y,r)}{\partial r}\right\}_{r=r(\alpha,\beta,I)} + O(\Delta^2(\epsilon))]dxdy \; , \quad \text{as } \epsilon \to 0$$

$$= I(\alpha,\beta,\rho) + \int_{\alpha}^{\alpha+\epsilon}\int_{-\infty}^{\beta} \psi(x,y,\rho)dxdy + \Delta(\epsilon)\int_{-\infty}^{\alpha+\epsilon}\int_{-\infty}^{\beta}\left\{\frac{\partial\psi(x,y,r)}{\partial r}\right\}_{r=\rho} dxdy + O(\Delta^2(\epsilon))$$

Hence

$$\frac{\Delta(\epsilon)}{\epsilon} = - \left[\frac{\partial I(a,b,r)}{\partial r}\right]^{-1}_{\alpha+\epsilon,\beta,\rho} \cdot \frac{1}{\epsilon} \int_{\alpha}^{\alpha+\epsilon} \int_{-\infty}^{\beta} \psi(x,y,\rho) + \frac{\Delta(\epsilon)}{\epsilon} \cdot o(\Delta(\epsilon)) \; .$$

Thus,

$$\left.\frac{\partial r(a,b,W)}{\partial a}\right|_{\alpha,\beta,I(\alpha,\beta,\rho)} = \lim_{\epsilon\to 0} \frac{\Delta(\epsilon)}{\epsilon}$$

$$= - \lim_{\epsilon\to 0} \left[\frac{\partial I(a,b,r)}{\partial r}\right]^{-1}_{\alpha+\epsilon,\beta,\rho} \cdot \lim_{\epsilon\to 0} \frac{1}{\epsilon} \int_{\alpha}^{\alpha+\epsilon} \int_{-\infty}^{\beta} \psi(x,y,\rho)dxdy$$

$$= - \left[\frac{\partial I(\alpha,\beta,\rho)}{\partial r}\right]^{-1} \cdot \varphi(\alpha)\ \Phi\left(\frac{\beta-\rho\alpha}{(1-\rho^2)^{1/2}}\right) \; .$$

Similarly,

$$\left.\frac{\partial r(a,b,W)}{\partial b}\right|_{\alpha,\beta,I(\alpha,\beta,\rho)}$$

$$= \left[\frac{\partial I(\alpha,\beta,\rho)}{\partial r}\right]^{-1} \cdot \varphi(\beta)\ \Phi\left(\frac{\alpha-\rho\beta}{(1-\rho^2)^{1/2}}\right) \; ,$$

and

$$\left.\frac{\partial r(a,b,W)}{\partial W}\right|_{\alpha,\beta,I(\alpha,\beta,\rho)} = \left[\frac{\partial I(\alpha,\beta,\rho)}{\partial r}\right]^{-1} \; .$$

It now follows from another application of Cramér's theorem that $\sqrt{n}\ (r_n^{(5)}-\rho)/h(\rho)$ is asymptotically normal with mean 0 and variance 1 , where

$$h^2(\rho) = [\frac{\partial I(\alpha,\beta,\rho)}{\partial r}]^{-2} \cdot [\Phi(\alpha)(1-\Phi(\alpha))\Phi^2(\frac{\beta-\rho\alpha}{\sqrt{1-\rho^2}}) + \Phi(\beta)(1-\Phi(\beta))\Phi^2(\frac{\alpha-\rho\beta}{\sqrt{1-\rho^2}})$$

$$+ I(\alpha,\beta,\rho)(1-I(\alpha,\beta,\rho))+2\Phi(\frac{\beta-\rho\alpha}{\sqrt{1-\rho^2}})\ \Phi(\frac{\alpha-\rho\beta}{\sqrt{1-\rho^2}})(I(\alpha,\beta,\rho)-\Phi(\alpha)\Phi(\beta))$$

$$- 2\Phi(\frac{\beta-\rho\alpha}{\sqrt{1-\rho^2}})(1-\Phi(\alpha))I(\alpha,\beta,\rho)-2\Phi(\frac{\alpha-\rho\beta}{\sqrt{1-\rho^2}})(1-\Phi(\beta))I(\alpha,\beta,\rho)] \quad .$$

When $\rho=0$, $I(\alpha,\beta,\rho) = \Phi(\alpha)\ \Phi(\beta)$, and

$$\frac{\partial I(\alpha,\beta,\rho)}{\partial r} = (\frac{\rho}{1-\rho^2})I(\alpha,\beta,\rho) + \int_{-\infty}^{\alpha}\int_{-\infty}^{\beta} [\frac{\rho}{(1-\rho^2)^2}(x^2+y^2) - \frac{1+\rho^2}{(1-\rho^2)^2}xy]\psi(x,y,\rho)dxdy$$

$$= - [\int_{-\infty}^{\alpha} x\varphi(x)dx]\cdot[\int_{-\infty}^{\beta} x\varphi(x)dx]$$

$$= - \varphi(\alpha)\ \varphi(\beta) \quad .$$

Hence

$$h^2(0) = \frac{\Phi(\alpha)(1-\Phi(\alpha))\Phi(\beta)(1-\Phi(\beta))}{\varphi^2(\alpha)\ \varphi^2(\beta)} \quad ,$$

and under the null hypothesis

$$\frac{\sqrt{n}\ r_n^{(5)}\ \ \varphi(\alpha)\ \varphi(\beta)}{[\Phi(\alpha)(1-\Phi(\alpha))\Phi(\beta)(1-\Phi(\beta))]^{1/2}}$$

is asymptotically normal with mean zero and variance 1. But this involves α and β which are unknown. However, since $\overline{U}_n$ and $\overline{V}_n$ are consistent estimates of $\Phi(\alpha)$ and $\Phi(\beta)$ respectively and a_n and b_n are consistent estimates of α and β respectively, the statistic

$$\frac{\sqrt{n}\, r_n^{(5)}\, \varphi(a_n)\, \varphi(b_n)}{[\bar{U}_n(1-\bar{U}_n)\bar{V}_n(1-\bar{V}_n)]^{1/2}}$$

is asymptotically normal with mean 0 and variance 1 under the null hypothesis. Thus the statistic

$$T_n^{(5)} = \frac{n\, r_n^{(5)^2}\, \varphi^2(a_n)\, \varphi^2(b_n)}{\bar{U}_n(1-\bar{U}_n)\bar{V}_n(1-\bar{V}_n)}$$

has χ^2 distribution with 1 d.f. under the null hypothesis. When $\rho \neq 0$,

$$\sqrt{n}\ \left[\frac{r_n^{(5)}\, \varphi(a_n)\, \varphi(b_n)}{[\bar{U}_n(1-\bar{U}_n)\bar{V}_n(1-\bar{V}_n)]^{1/2}} - \frac{\rho\, \varphi(\alpha)\, \varphi(\beta)}{[\Phi(\alpha)(1-\Phi(\alpha))\Phi(\beta)(1-\Phi(\beta))]^{1/2}}\right]$$

is asymptotically normal with mean 0 and some variance which we do not need.

Most of the asymptotic distributions discussed in this section are known and they are given here only for the sake of completeness. We summarize these results below.

$$\sqrt{n}\ \left(\frac{r_n^{(1)}}{\sqrt{1-r_n^{(1)^2}}} - \frac{\rho}{\sqrt{1-\rho^2}}\right)$$

is asymptotically normal with mean 0 and variance $h_1^2(\rho)$ where $h_1^2(0) = 1$. Thus when $\rho = 0$, the statistic

$$T_n^{(1)} = \frac{n\, r_n^{(1)^2}}{1-r_n^{(1)^2}}$$

is asymptotically χ^2 with 1 d.f., and when $\rho \neq 0$, $\frac{T_n^{(1)}}{n}$ is a consistent estimate of $\rho^2/(1-\rho^2)$.

$$\sqrt{n} \ (r_n^{(3)} + \frac{\rho \ \varphi(\beta)}{[\Phi(\beta)(1-\Phi(\beta))]^{1/2}}) ,$$

is asymptotically normal with mean 0 and variance $h_3^2(\rho)$ where $h_3^2(0) = 1$. Thus, when $\rho = 0$, the statistic

$$T_n^{(3)} = nr_n^{(3)^2}$$

is asymptotically χ^2 with 1 d.f., and when $\rho \neq 0$, $\frac{T_n^{(3)}}{n}$ is a consistent estimate of $\rho^2\varphi^2(\beta)/[\Phi(\beta)(1-\Phi(\beta))]$.

If we try to derive a statistic from $r_n^{(2)}$ whose asymptotic null distribution is completely known, we get the statistic $T_n^{(3)}$ which is being considered in connection with $r_n^{(3)}$. So we shall not discuss any more about tests of independence based on $r_n^{(2)}$.

$$\sqrt{n} \ (\frac{r_n^{(4)}-\rho}{\sqrt{1.939}})$$

is asymptotically normal with mean zero and variance $h_4^2(\rho)$ where $h_4^2(0) = 1$. Thus when $\rho = 0$, the statistic

$$T_n^{(4)} = \frac{n \ r_n^{(4)^2}}{1.939}$$

is asymptotically χ^2 with 1 d.f., and when $\rho \neq 0$, $\frac{T_n^{(4)}}{n}$ is a consistent estimate of $\rho^2/1.939$. $T_n^{(4)}$ can be used only in the special case when $\nu = \eta$ and ξ and σ^2 are known.

$$\sqrt{n}\left(\frac{r_n^{(5)}\ \varphi(a_n)\ \varphi(b_n)}{[\bar{U}_n(1-\bar{U}_n)\bar{V}_n(1-\bar{V}_n)]^{1/2}} - \frac{\rho\ \varphi(\alpha)\ \varphi(\beta)}{[\Phi(\alpha)(1-\Phi(\alpha))\Phi(\beta)(1-\Phi(\beta))]^{1/2}}\right)$$

is asymptotically normal with mean 0 and variance $h_5^2(\rho)$ where $h_5^2(0) = 1$. Thus when $\rho = 0$, the statistic

$$T_n^{(5)} = \frac{n\ r_n^{(5)^2}\ \varphi^2(a_n)\ \varphi^2(b_n)}{\bar{U}_n(1-\bar{U}_n)\bar{V}_n(1-\bar{V}_n)}$$

is asymptotically χ^2 with 1 d.f., and when $\rho \neq 0$, $\frac{T_n^{(5)}}{n}$ is a consistent estimate of

$$\frac{\rho^2\ \varphi^2(\alpha)\ \varphi^2(\beta)}{\Phi(\alpha)(1-\Phi(\alpha))\Phi(\beta)(1-\Phi(\beta))} .$$

$$\sqrt{n}\left(r_n^{(6)} - \frac{I(\alpha,\beta,\rho)-\Phi(\alpha)\Phi(\beta)}{[\Phi(\alpha)(1-\Phi(\alpha))\Phi(\beta)(1-\Phi(\beta))]^{1/2}}\right)$$

is asymptotically normal with mean 0 and variance $h_6^2(\rho)$ where $h_6^2(0) = 1$. Thus the statistic

$$T_n^{(6)} = nr_n^{(6)^2} ,$$

which is the statistic used for testing independence in a 2 x 2 contingency table, is asymptotically χ^2 with 1 d.f., and when $\rho \neq 0$, $\frac{T_n^{(6)}}{n}$ is a consistent estimate of

$$\frac{[I(\alpha,\beta,\rho)-\Phi(\alpha)\ \Phi(\beta)]^2}{\Phi(\alpha)\ (1-\Phi(\alpha))\Phi(\beta)(1-\Phi(\beta))} .$$

Each of the statistics $T_n^{(1)}$, $T_n^{(3)}$, $T_n^{(4)}$, $T_n^{(5)}$ and $T_n^{(6)}$ is asymptotically χ^2 with 1 d.f. under the null hypothesis $H_o:\rho=0$, and each of them approaches ∞ in probability when $\rho \neq 0$. Thus on the basis of $T_n^{(i)}$ we can have the following test for H_o:

Test i: Reject H_o if $T_n^{(i)} > c_\alpha$,

where c_α is such that $P[\chi_1^2 > c_\alpha] = \alpha$, and achieve the level of significance α in the limit, or use the statistic

$$P[\chi_1^2 \leq T_n^{(i)}]$$

as a basis for judging how strongly a departure from H_o is indicated by the data.

4. Asymptotic properties of tests considered in section 3.

Pitman-efficiencies of tests 3, 4, 5, 6 relative to test 1:

Suppose $Z_1,\ldots,Z_n,\ldots$ is a sequence of random variables whose distributions depend on a parameter θ. Suppose further that the limiting distribution of $\sqrt{n}\,(Z_n-f(\theta))/g(\theta)$ is normal with mean 0 and variance 1 where $f(0) = 0$, and $g(\theta)$ is continuous at $\theta = 0$ with $g(0) = 1$. Consider the test for the hypothesis H_o: $\theta=0$, based on Z_n which rejects H_o if $\sqrt{n}\,|Z_n| > c$ where $c = \Phi^{-1}(1 - \frac{\alpha}{2})$. This sequence of tests has a limiting level of significance α. Let $P_\theta[\sqrt{n}|Z_n| > c]$ indicate the probability of the event $\sqrt{n}|Z_n| > c$ when θ obtains. Then $P_\theta[\sqrt{n}|Z_n| > c]$ is the power of the above test based on Z_n at θ. Now consider a sequence of alternatives, $\theta_n = \delta/\sqrt{n}$. It can be shown that

$$\lim_{n\to\infty} P_{\theta_n}[\sqrt{n}|Z_n| > c]$$

$$= 1 - [\Phi(\Phi^{-1}(1 - \frac{\alpha}{2})+\delta\cdot f'(0))-\Phi(-\Phi^{-1}(1 - \frac{\alpha}{2})+\delta\cdot f'(0))] \quad ,$$

where $f'(0)$ is the first derivative of the function f at $\theta = 0$. This is the limiting power of the sequence of tests based on $\{Z_n\}$, fixing our attention to the sequence of alternatives $\{\frac{\delta}{\sqrt{n}}\}$. (For a proof of the above statement, see Noether [8]). It can be easily seen that this limiting power increases as $|\delta f'(0)|$ increases. From the above formula for limiting power, it can be concluded that if there are two sequences of tests based on two sequences of random variables, each having the same limiting level of significance and if these two sequences or random variables satisfy the above conditions with $f_1(\theta)$ and $f_2(\theta)$, then in order that the two tests attain the same limiting power in the above sense at the same value of θ (of the order of $n^{-1/2}$), the limiting ratio of the sample size needed by the first test to the sample size needed by the second test is $[f_2'(0)/f_1'(0)]^2$. This is known as the Pitman-efficiency of the second test relative to the first test.

Each of the five tests mentioned at the end of section 3 correspond to the above set up with ρ playing the role of θ. If we denote by f_i, the function corresponding to test i, then

$$f_1(\rho) = \frac{\rho}{\sqrt{1-\rho^2}} \;, \; f_3(\rho) = -\frac{\rho\,\varphi(\beta)}{[\Phi(\beta)(1-\Phi(\beta))]^{1/2}} \;, \quad f_4(\rho) = \frac{\rho}{\sqrt{1.939}} \;,$$

$$f_5(\rho) = \frac{\rho\,\varphi(\alpha)\,\varphi(\beta)}{[\Phi(\alpha)(1-\Phi(\alpha))\Phi(\beta)(1-\Phi(\beta))]^{1/2}} \;, \text{ and}$$

$$f_6(\rho) = \frac{I(\alpha,\beta,\rho)-\Phi(\alpha)\,\Phi(\beta)}{[\Phi(\alpha)(1-\Phi(\alpha))\Phi(\beta)(1-\Phi(\beta))]^{1/2}} \;.$$

Hence

$$f_1'(0) = 1, \; f_3'(0) = -\frac{\varphi(\beta)}{[\Phi(\beta)(1-\Phi(\beta))]^{1/2}} \;, \quad f_4'(0) = \frac{1}{\sqrt{1.939}} \;,$$

$$f_5'(0) = \frac{\varphi(\alpha)\;\;\varphi(\beta)}{[\Phi(\alpha)(1-\Phi(\alpha))\Phi(\beta)(1-\Phi(\beta))]^{1/2}} \;, \text{ and}$$

$$f_6'(0) = -\frac{\varphi(\alpha)\,\varphi(\beta)}{[\Phi(\alpha)(1-\Phi(\alpha))\Phi(\beta)(1-\Phi(\beta))]^{1/2}} \;.$$

If we denote the Pitman-efficiency of test i relative to test 1 by E^*_{i1}, then we have

$$E^*_{31} = \frac{\varphi^2(\beta)}{\Phi(\beta)(1-\Phi(\beta))} \;,$$

$$E^*_{41} = \frac{1}{1.939} = 0.5157 \;,$$

and

$$E^*_{51} = E^*_{61} = \frac{\varphi^2(\alpha)\;\;\varphi^2(\beta)}{\Phi(\alpha)(1-\Phi(\alpha))\Phi(\beta)(1-\Phi(\beta))} \;.$$

Bahadur - efficiencies of tests 3, 4, 5, 6 relative to test 1:

It has already been mentioned at the end of section 3 that one can use the statistic

$$P[X_1^2 \le T_n^{(i)}]$$

as a basis for judging how strongly a departure from $H_o:\rho=0$ is indicated by the data.

Now when $\rho=0$, $T_n^{(i)}$ is asymptotically X^2 with 1 d.f. Hence the statistic

$$K_n^{(i)} = -2\log P[X_1^2 > T_n^{(i)}] ,$$

which is an increasing function of $P[X_1^2 \le T_n^{(i)}]$, is distributed as X^2 with 2 d.f., and we see that $\frac{K_n^{(i)}}{n}$ converges to 0 in probability.

When $\rho \ne 0$, we have already seen that $\frac{T_n^{(i)}}{n}$ is a consistent estimate of some positive function of ρ, say $b_i(\rho)$. Thus $T_n^{(i)}$ tends to ∞ in probability as $n \to \infty$ when $\rho \ne 0$. But

$$-2 \log P[X_1^2 > x] = -2 \log P[X_1 > \sqrt{x}] = x[1+o(1)], \text{ as } x \to \infty .$$

Hence when $\rho \ne 0$ and n is large

$$\begin{aligned} K_n^{(i)}/n &= -2 \log P[X_1^2 > T_n^{(i)}]/n \\ &\simeq T_n^{(i)}/n \end{aligned}$$

with very high probability and $\frac{T_n^{(i)}}{n}$ converges in probability to $b_i(\rho)$. Thus when $\rho \ne 0$, $\frac{K_n^{(i)}}{n}$ converges to $b_i(\rho) > 0$ in probability.

Since $K_n^{(i)}$ gives us an idea about how good the statistic $T_n^{(i)}$ is for detecting a deviation from the null hypothesis, Bahadur [1] has suggested the possibility of using $b_i(\rho)$ as a measure of efficacy of the test based on $\{T_n^{(i)}\}$, and calls it the asymptotic slope of the sequence of tests based on $\{T_n^{(i)}\}$. If there are two subsequences of integers $\{n_r^{(i)}\}$ and $\{n_r^{(j)}\}$ such that

$$K^{(i)}_{n_r^{(i)}} \,/\, K^{(j)}_{n_r^{(j)}}$$

tends to one in probability as $r \to \infty$, then for large r the tests based on $T^{(i)}_{n_r^{(i)}}$ and $T^{(j)}_{n_r^{(j)}}$ are approximately equally effective to test the null hypothesis. For this reason $\{n_r^{(i)}\}$ and $\{n_r^{(j)}\}$ satisfying the above condition, are called asymptotically equivalent sample sizes. It follows that $\{n_r^{(i)}\}$ and $\{n_r^{(j)}\}$ are asymptotically equivalent sample sizes for tests based on $\{T_n^{(i)}\}$ and $\{T_n^{(j)}\}$ respectively, if and only if

$$\lim_{r\to\infty} \frac{n_r^{(i)}}{n_r^{(j)}} = \frac{b_j(\rho)}{b_i(\rho)} .$$

For this reason Bahadur calls $b_j(\rho)/b_i(\rho)$, the asymptotic efficiency of the sequence of tests based on $\{T_n^{(j)}\}$ relative to the sequence of tests based on $\{T_n^{(i)}\}$.

We have already seen that

$$b_1(\rho) = \frac{\rho^2}{1-\rho^2} \;,\quad b_3(\rho) = \frac{\rho^2\,\varphi^2(\beta)}{\Phi(\beta)(1-\Phi(\beta))} \,,\quad b_4(\rho) = \frac{\rho^2}{1.939} \;,$$

$$b_5(\rho) = \frac{\rho^2\varphi^2(\alpha)\,\varphi^2(\beta)}{\Phi(\alpha)(1-\Phi(\alpha))\Phi(\beta)(1-\Phi(\beta))} \,, \text{ and } b_6(\rho) = \frac{[I(\alpha,\beta,\rho)-\Phi(\alpha)\Phi(\beta)]^2}{\Phi(\alpha)(1-\Phi(\alpha))\Phi(\beta)(1-\Phi(\beta))} .$$

If we denote the Bahadur-efficiency of the sequence of tests based on $\{T_n^{(i)}\}$ relative to the sequence of tests based on $\{T_n^{(1)}\}$ by E_{i1}, then we have

$$E_{31} = \frac{\varphi^2(\beta)}{\Phi(\beta)(1-\Phi(\beta))} \cdot (1-\rho^2) = (1-\rho^2)\, E^*_{31} ,$$

$$E_{41} = \frac{1-\rho^2}{1.939} = 0.5157(1-\rho^2) = (1-\rho^2)\, E^*_{41} ,$$

$$E_{51} = \frac{\varphi^2(\alpha)\, \varphi^2(\beta)}{\Phi(\alpha)(1-\Phi(\alpha))\Phi(\beta)(1-\Phi(\beta))} \cdot (1-\rho^2) = (1-\rho^2)\, E^*_{51}$$

and $$E_{61} = \frac{[I(\alpha,\beta,\rho)-\Phi(\alpha)\Phi(\beta)]^2}{\Phi(\alpha)(1-\Phi(\alpha))\Phi(\beta)(1-\Phi(\beta))} \cdot \frac{1-\rho^2}{\rho^2}$$

5. Tables and conclusions.

In this section we shall present some tables for the Pitman-efficiencies and Bahadur efficiencies of tests 3, 4, 5 and 6 relative to test 1. These tables are prepared from the formulas given at the end of section 4.

TABLE 1

ASYMPTOTIC EFFICIENCIES OF TESTS BASED ON THE POINT BISERIAL (Test 3) AND FLANAGAN (Test 4) CORRELATION COEFFICIENTS RELATIVE TO TEST BASED ON MAXIMUM LIKELIHOOD ESTIMATE (Test 1)

		Pitman-efficiency	Bahadur-efficiency		
			$\rho = \pm .10$	$\rho = \pm .25$	$\rho = \pm .50$
Test 3	$\beta = 0$	.637	.630	.597	.477
	$\beta = \pm 0.5$	.581	.575	.545	.436
	$\beta = \pm 1.0$	.439	.434	.411	.329
	$\beta = \pm 1.5$	.269	.266	.252	.202
	$\beta = \pm 1.0$	.131	.130	.123	.098
Test 4		.515	.510	.484	.386

TABLE 2

ASYMPTOTIC EFFICIENCIES OF TESTS BASED ON TETRACHORIC (Test 5) AND PHI(Test 6) CORRELATION COEFFICIENTS RELATIVE TO TEST BASED ON MAXIMUM LIKELIHOOD ESTIMATE (Test 1)

		Pitman-efficiency	Bahadur-efficiency					
			$\rho=-.50$	$\rho=-.25$	$\rho=-.10$	$\rho=+.10$	$\rho=+.25$	$\rho=+.50$
$\alpha=0, \beta=0$	Test 5	.405	.304	.380	.401	.401	.380	.304
	Test 6	.405	.333	.388	.403	.403	.388	.333
$\alpha=0, \beta=0.5$	Test 5	.370	.277	.347	.366	.366	.347	.277
	Test 6	.370	.297	.352	.367	.367	.352	.297
$\alpha=0, \beta=1.0$	Test 5	.279	.209	.262	.276	.276	.262	.209
	Test 6	.279	.208	.262	.276	.276	.262	.208
$\alpha=0, \beta=1.5$	Test 5	.171	.128	.161	.170	.170	.161	.128
	Test 6	.171	.113	.156	.169	.169	.156	.113
$\alpha=0, \beta=2.0$	Test 5	.083	.063	.078	.083	.083	.078	.063
	Test 6	.083	.047	.073	.082	.082	.073	.047
$\alpha=0.5, \beta=0.5$	Test 5	.338	.253	.316	.334	.334	.316	.253
	Test 6	.338	.229	.300	.326	.343	.341	.306
$\alpha=0.5, \beta=1.0$	Test 5	.255	.191	.239	.252	.252	.239	.191
	Test 6	.255	.140	.209	.240	.265	.562	.248
$\alpha=0.5, \beta=1.5$	Test 5	.156	.117	.147	.155	.155	.147	.117
	Test 6	.156	.068	.117	.143	.166	.172	.155
$\alpha=0.5, \beta=2.0$	Test 5	.076	.057	.071	.075	.075	.071	.057
	Test 6	.076	.026	.052	.068	.083	.086	.073
$\alpha=1.0, \beta=1.0$	Test 5	.192	.144	.180	.190	.190	.180	.144
	Test 6	.192	.077	.137	.172	.210	.229	.235
$\alpha=1.0, \beta=1.5$	Test 5	.118	.089	.111	.117	.117	.111	.089
	Test 6	.118	.034	.073	.100	.135	.156	.172
$\alpha=1.0, \beta=2.0$	Test 5	.058	.043	.054	.057	.057	.054	.043
	Test 6	.058	.012	.031	.046	.069	.084	.094
$\alpha=1.5, \beta=1.5$	Test 5	.072	.054	.068	.072	.072	.068	.054
	Test 6	.072	.014	.037	.056	.089	.115	.148
$\alpha=1.5, \beta=2.0$	Test 5	.035	.026	.033	.035	.035	.033	.026
	Test 6	.035	.005	.015	.026	.047	.066	.097
$\alpha=2.0, \beta=2.0$	Test 5	.017	.013	.016	.017	.017	.016	.013
	Test 6	.017	.002	.005	.011	.025	.041	.076

Note. In table 2, efficiencies are given for $\beta \geq \alpha \geq 0$. The following formulas will give efficiencies for negative values of α, β and for $\alpha \geq \beta \geq 0$.

$$E^*_{51}(\alpha,\beta) = E^*_{51}(\beta,\alpha) = E^*_{51}(\alpha,-\beta) = E^*_{51}(-\alpha,\beta) \quad ,$$

$$E_{51}(\alpha,\beta,\rho) = E_{51}(\alpha,\beta,-\rho) = E_{51}(\beta,\alpha,\rho) = E_{51}(\alpha,-\beta,\rho) = E_{51}(-\alpha,\beta,\rho) \quad ,$$

$$E^*_{61}(\alpha,\beta) = E^*_{61}(\beta,\alpha) = E^*_{61}(\alpha,-\beta) = E^*_{61}(-\alpha,\beta) \quad ,$$

$$E_{61}(\alpha,-\beta,\rho) = E_{61}(-\alpha,\beta,\rho) = E_{61}(\alpha,\beta,-\rho) \quad ,$$

$$E_{61}(\alpha,\beta,\rho) = E_{61}(\beta,\alpha,\rho) = E_{61}(-\alpha,-\beta,\rho) \quad .$$

TABLE 3

ASYMPTOTIC EFFICIENCIES OF POINT BISERIAL (Test 3), FLANAGAN (Test 4), TETRACHORIC (Test 5) AND PHI (Test 6) CORRELATION COEFFICIENTS RELATIVE TO TEST BASED ON MAXIMUM LIKELIHOOD ESTIMATE (Test 1) FOR OPTIMUM DICHOTOMY $(\alpha = 0, \ \beta = 0)$

	Pitman-efficiency	Bahadur-efficiency		
		$\rho = \pm .10$	$\rho = \pm .25$	$\rho = \pm .50$
Test 3	.637	.630	.597	.472
Test 4	.515	.510	.484	.386
Test 5	.405	.401	.380	.304
Test 6	.405	.403	.388	.333

We shall now try to get some answers to the questions posed in section 1 in the light of the Pitman-efficiencies and Bahadur-efficiencies for the different tests. From the tables given above we conclude that

(i) The optimum dividing point for defining the dichotomy on a variate is the mean of that variate. These means are unknown in our formulation of the problem, but in many cases the experimenter has prior information about the means and this information can be used to an advantage.

(ii) When optimum dichotomy is used on a single variate, the test based on point biserial correlation is asymptotically more efficient than the test suggested by Mosteller. The above comparison remains valid even when the dividing point is one-half standard deviation off the mean of the variate under question.

(iii) When optimum dichotomy is used on both variates, the test 6 for independence in 2x2 contingency table (which does not make any use of the knowledge of bivariate normality) is asymptotically more efficient in the sense of Bahadur than the test 5 based on tetrachoric correlation. However, the two tests becomes more and more alike as the value of $\rho \neq 0$ to be detected, becomes closer and closer to 0. Finally, they attain the same Pitman-efficiency relative to test 1. For non-optimal dichotomies, the Bahadur-efficiency of test 6 goes down at a much sharper rate than the Bahadur-efficiency of test 5 if either only one or all three of the quantities α, β and ρ are negative. When either none or two of the three quantities α, β and ρ

are negative, test 5 becomes more efficient than test 6 only when one dichotomy is very close to optimal while the other is far from optimal.

(iv) An optimal dichotomy used on a single variate results in about 36% loss of information in detecting small departures from the null hypothesis. For larger departures from the null hypothesis, the loss of information is still higher.

(v) Optimal dichotomy used on both variates results in about 60% loss of information in detecting small departures from the null hypothesis. For larger departures from the null hypothesis, the loss of information is still higher.

6. Acknowledgements.

The author is indebted to Herbert Solomon for many helpful suggestions in preparing the paper. Thanks are also due to Phyllis Groll for preparing the tables given in section 4.

References

[1] R. R. Bahadur, "Stochastic comparison of tests", Ann. Math. Stat. (1960), vol. 31, pp. 276-295.

[2] H. Cramér, "Mathematical methods of statistics", Princeton University Press (1946), pp. 366-367.

[3] J. C. Flanagan, "General considerations in the selection of test items and a short method of estimating the product-moment coefficient from the tails of the distribution", Jour. of Educ. Psych. (1939), vol. 30, pp. 674-680.

[4] T. L. Kelley, "The selection of upper and lower groups for the validation of test items", Jour. of Educ. Psych. (1939), vol. 30, pp. 17-24.

[5] M. G. Kendall, "The advanced theory of statistics", vol. 1, 3rd ed. Charles Griffin Co. (1947), pp. 355-356.

[6] E. L. Lehmann, "Testing statistical hypotheses", John Wiley & Sons, Inc. (1959), p. 251.

[7] F. Mosteller, "On some useful inefficient statistics", Ann. Math. Stat. (1946), vol. 17, pp. 377-408.

[8] G. E. Noether, "A theorem of Pitman", Ann. Math. Stat. (1955), vol. 26, pp. 64-68.

[9] K. Pearson, "On the correlation of characters not quantitatively measurable", Phil. Trans. Series A (1901), vol. 195, pp. 1-47.

[10] K. Pearson, "On a new method for determining the correlation between a measured character A and a measured character B of which only the percentage of cases wherein B exceeds (or falls short of) a given intensity is recorded for each grade of A". Biometrika (1909), vol. 7, pp. 96-105.

[11] R. F. Tate, "Correlation between a discrete and a continuous variable", Ann. Math. Stat. (1954), vol. 25, pp. 603-607.

THE SCORING OF MULTIPLE CHOICE QUESTIONNAIRES

by

Herman Chernoff

1. Summary.

In many types of questionnaires the element of guessing is important. An approach to correcting for guessing is proposed. Here one regards the score assigned to a subject on a question as an estimate of the unknown value of the subject's knowledge on the question. This value is one when the subject knows the answer and is zero otherwise. To derive scores with minimum mean squared error, it is necessary to consider the responses of the whole population of subjects. Thus the score for a correct answer to a question depends on the proportion p of correct answers in the population. In the simplest model, we assume that a subject who knows the answer, responds correctly and that others select a response at random among r choices. Then an incorrect response is scored zero and a correct one is assigned a score of λ/p where λ is the proportion of the population who knew the answer and can be estimated using the relation $p = \lambda + (1-\lambda)/r$.

The general approach is also applied to a variety of more complicated models, each of which are examples of a specified general formulation. These models include the "pairs of questions' model the "partial knowledge model and the "scaled questions" model.

While the method applies neatly to single questions, there are fundamental difficulties in extending it to obtain an overall or composite score for a subject on an examination consisting of many questions. This problem is discussed briefly and the simple minded procedure of totaling the scores for the individual questions is partially evaluated.

2. Introduction.

We indicate a general approach to the scoring of examinations with multiple choice questions. The main idea may be motivated by referring to a test question which offers three choices, A, B, and C and for which we can assume that A is the correct answer. Suppose that when the results for all of the subjects are tabulated, the three choices A, B, and C have been selected with equal frequencies. If there is a large number of subjects, these frequencies indicate that at most a negligible portion of the subjects knew the answer. The customary procedure of scoring a plus one when a subject chooses A would be misleading. It would constitute the throwing of an irrelevant random "noise" into the total test score. Clearly, giving each subject a zero or throwing out the question altogether would be preferable. Suppose, on the other hand, that when the results are tabulated, A is selected 90% of the time, and B and C, each 5% of the time. Then it would seem reasonable to assume that 85% of the subjects knew the answer and only 5% guessed it correctly. Here it would make much more sense to score a plus one for each subject who selected A .

Comparing the two cases above, it seems reasonable to expect that a proper procedure for scoring an individual should depend not only on the subject's choice, but on the frequencies of the choices for the entire population of subjects. Once this is granted, it remains to decide on an optimal scoring procedure. The general approach, which consists of regarding the score as an estimate whose mean squared error should be minimized, is idscussed and is illustrated in section 3 for the simplest model. In section 4, a generalization where two questions are considered simultaneously is discussed. Then the method is generalized in section 5.

There it is seen that the necessary calculus is essentially that of conditional expectation or regression.

Sections 6, 7, and 8 are devoted to other variations of the model. Finally, in section 9, the question of overall or composite scores is discussed and miscellaneous remarks are left for section 10.

3. The simplest model.

In the simplest model we assume that there are r possible choices for a question. A subject who knows the answer gives the correct answer. One who doesn't know the answer guesses and has probability $1/r$ of arriving at the correct response.

Suppose that the proportion of the population who know the answer is λ . The proportion who answer correctly is given by

$$p = \lambda + \frac{1}{r}(1-\lambda) \quad . \tag{3.1}$$

For a large sample of respondents p and λ may be estimated accurately. Suppose that each individual who obtains the correct answer receives a score of $x = x_c$ and the others receive a score of $x = x_w$. We regard x as an estimate of the appropriate value which is one for an individual who knows the answer and zero for the others. The respondents can be partitioned into three groups. There is a proportion λ who know the answer and respond correctly. For them the error of estimate is $x_c - 1$. There is a second group, a proportion $p-\lambda$, who do not know the answer but guess correctly. For them the error of estimate is x_c since they deserve zero. Finally there is the proportion $1-p$ who do not know the answer and who guess incorrectly. For them the error of estimate is x_w. Consequently the mean squared error is given by

$$V = \lambda(x_c-1)^2 + (p-\lambda)\, x_c^2 + (1-p)\, x_w^2 \quad .$$

Minimizing with respect to x_c and x_w we obtain

$$(3.2) \qquad \begin{aligned} x_c &= \frac{\lambda}{p} = \frac{rp-1}{(r-1)p} \\ x_w &= 0 \end{aligned}$$

giving a minimum value of

$$(3.3) \qquad V = \frac{\lambda(p-\lambda)}{p} \, .$$

This compares favorably with

$$(3.4) \qquad V^* = p-\lambda$$

which is the mean squared error for the usual scoring system which assigns a one to a correct answer and a zero to an incorrect answer.

The scoring method represented by equations (3.2) is subject to two criticisms which we shall now consider. First, one may remark that the scores obtained by this technique are perfectly correlated with those obtained by the usual method. Thus nothing is gained by using this new approach. The second criticism is that a subject's scores will depend not only on his response but also on the responses of the other subjects. In fact a subject who knows the answers to each question on an examination will receive very little credit if most of the other subjects are ignorant. This criticism was leveled by Lyerly [10] at a technique formerly proposed by Hamilton [6] to correct for guessing.

The first criticism oversimplifies the situation. If an examination consists of several questions the sum of a subject's scores under the proposed system will not correlate perfectly with the sum of the scores under the usual system. There is one exception to the last remark. That is the case where all the questions are equally difficult to the population of subjects in the sense that the corresponding λ are equal. Even in this case a modification that we shall soon propose will invalidate the first criticism.

The second criticism may be serious in those cases where it is desired to compare subjects with an objective standard (as expressed by the examination) and an underestimate of the subject's knowledge is worse than an overestimate. In other cases the criticism may be disregarded. If an underestimate is no worse than an overestimate, the proposed system is superior. If it is merely desired to rank the subjects, the proposed system is adequate. If it is merely desired to give the subject who answers all the questions correctly a perfect score, one may compare the total scores with the one that would be given to a subject with all answers correct.

Finally, the following modification will eliminate the first criticism and reduce the effect of the second when it is applicable. If the population of subjects is large, one may subdivide them or stratify them into subgroups and score each subgroup separately. If the stratification is performed according to some criterion which correlates with ability to answer the questions, the new procedure would have improved performance.

For example suppose that the population is divided into three groups of equal size. In the first almost no one knows the answer, in the second half of the subjects know the answer, and in the third almost all of the subjects know the answer. The scores for correct answers in these subgroups would be 0, $(1+r^{-1})^{-1}$, and 1 respectively. The overall mean squared error would be $[6(r+1)]^{-1}$ compared with $[2(r+1)]^{-1}$ if the population is not stratified.

If stratification is applied, the scores are no longer perfectly correlated with those of the usual procedure even when the examination consists of one question. Soundly applied stratification reduces the effect of the second criticism when it applies. One may note however that stratification could lead to unhappy individual consequences if a few very capable subjects were, by chance, included in the wrong stratum. Even this possibility can be avoided by variations of the following expedient. Obtain total scores without using stratification. Then use these total scores to stratify the population and recalculate new improved scores.

The last suggestion has the "advantage" of objectivity. The final score depends only on the responses of the subject and the population and not on any other extraneous criteria. One can elaborate on it. First one could estimate a regression of the λ for an individual item or question on the total scores. Second one could break up the examination into relevant subportions and then apply a multiple regression. It may even be feasible to apply some form of factor analysis. Finally one could iterate and use the new improved scores to stratify and then to compute further improved scores.

4. Pairs of questions.

In this section we shall examine a generalization of the simplest model where the scores on a question depend on the responses to a pair of questions. This model will serve to lay the groundwork for a proper generalization of the method applied in the preceding section. It also represents one attack on the problem of scoring responses so as to reduce the effect of the ignorance of some subjects on the scores of other subjects.

Let λ_{11} be the proportion of the subjects who know the answer to both questions, λ_{10} be the proportion who know the answer to the first only, λ_{01} be the proportion who know the answer to the second only, and λ_{00} be the proportion who don't know either answer. Similarly we define p_{cc}, p_{cw}, p_{wc}, and p_{ww} in terms of the proportions who answer the questions correctly (c) and incorrectly (w). We are assuming that subjects who don't know the answer select a response at random.

Then,

$$p_{cc} = \lambda_{11} + \frac{\lambda_{10}}{r} + \frac{\lambda_{01}}{r} + \frac{\lambda_{00}}{r^2} \quad , \tag{4.1}$$

$$p_{cw} = \lambda_{10}\frac{(r-1)}{r} + \lambda_{00}\frac{(r-1)}{r^2} \quad ,$$

$$p_{wc} = \lambda_{01}\frac{(r-1)}{r} + \lambda_{00}\frac{(r-1)}{r^2} \quad ,$$

$$p_{ww} = \lambda_{00}\left(\frac{r-1}{r}\right)^2 \quad .$$

Let x_{icc}, x_{icw}, x_{iwc}, x_{iww} be the scores assigned for the i^{th} question to individuals whose results for the pair of questions are cc, cw, wc, and ww respectively, $i = 1, 2$. Then, the minimum mean squared error of the estimate x_{icc} from the appropriate value (which is one for individuals who know the answer to the i$^{\text{th}}$ question and 0 for the others) is obtained as follows:

Each subject belongs to one of nine nonempty classes determined by his knowledge and his answers. To each is assigned a value due to the knowledge and a score due to the pair of answers. Minimizing the mean squared error consists of minimizing

$$V_1 = (\lambda_{11} + \frac{\lambda_{10}}{r})(x_{1cc} - 1)^2 + (\frac{\lambda_{01}}{r} + \frac{\lambda_{00}}{r^2})\, x_{1cc}^2 + \frac{(r-1)\lambda_{10}}{r}\,(x_{1cw} - 1)^2$$

$$+ \lambda_{00}\,\frac{(r-1)}{r^2}\,x_{1cw}^2 + [\lambda_{01}\,\frac{(r-1)}{r} + \lambda_{00}\,\frac{(r-1)}{r^2}]\,x_{1wc}^2 + \lambda_{00}(\frac{r-1}{r})^2\,x_{1ww}^2$$

and

$$V_2 = (\lambda_{11} + \frac{\lambda_{01}}{r})(x_{2cc} - 1)^2 + (\frac{\lambda_{10}}{r} + \frac{\lambda_{00}}{r^2})\, x_{2cc}^2 + \frac{r-1}{r}\,\lambda_{01}\,(x_{2cw} - 1)^2$$

$$+ \lambda_{00}\,\frac{(r-1)}{r^2}\,x_{2cw}^2 + [\lambda_{10}\,\frac{r-1}{r} + \lambda_{00}\,\frac{r-1}{r^2}]\,x_{2wc}^2 + \lambda_{00}\,(\frac{r-1}{r})^2\,x_{2ww}^2 \,.$$

We obtain

$$\begin{aligned}
x_{1cc} &= \frac{\lambda_{11} + \frac{\lambda_{10}}{r}}{p_{cc}} & x_{1wc} &= 0 \\
x_{1cw} &= \frac{(r-1)\,\lambda_{10}/r}{p_{cw}} & x_{1ww} &= 0 \\
x_{2cc} &= \frac{\lambda_{11} + \frac{\lambda_{01}}{r}}{p_{cc}} & x_{2wc} &= \frac{(r-1)\,\lambda_{01}/r}{p_{wc}} \\
x_{2cw} &= 0 & x_{2ww} &= 0
\end{aligned} \qquad (4.2)$$

which yield minimizing values of

$$V_1 = \frac{\left[\lambda_{11} + \frac{\lambda_{10}}{r}\right]\left[\frac{\lambda_{01}}{r} + \frac{\lambda_{00}}{r^2}\right]}{p_{cc}} + \frac{\left[\frac{(r-1)\ \lambda_{10}}{r}\right]\left[\frac{\lambda_{00}\ (r-1)}{r^2}\right]}{p_{cw}} ,$$

$$(4.3) \qquad V_2 = \frac{\left[\lambda_{11} + \frac{\lambda_{01}}{r}\right]\left[\frac{\lambda_{10}}{r} + \frac{\lambda_{00}}{r^2}\right]}{p_{cc}} + \frac{\left[\frac{(r-1)}{r}\lambda_{01}\right]\left[\frac{\lambda_{00}(r-1)}{r^2}\right]}{p_{cw}} .$$

Thus by combining two questions, one may obtain some reduction in the mean squared error for each question. Technically this reduction is obtained by using more information; namely the combined scores. To accomplish this requires more computation and an extension of the model. This extension involves assuming independence of the guesses for individuals who don't know the answers.

For an examination consisting of only two questions the method proposed in this section is an improvement over the suggestion in section 3 of stratifying by total scores. Its efficacy depends on the extent to which the knowledge of the answers to the two question are correlated.

In principle the model and results of this section can be extended to examinations of arbitrary length. They would give results superior to those based on stratification using only criteria involving responses on the examination. Unfortunately this extension is not very practical. The computational difficulties grow enormously as the number of questions increases. Furthermore the number of subjects necessary to estimate reliably the appropriate λ's also grows enormously. As a result, the author would recommend using this model or an extension of it only on a portion of the examination consisting of a small number of strongly "related" questions. The word related is used in the sense that knowledge of the answers are correlated.

Suppose that the two questions are of unequal difficulty but of equal value. Then the subject who answers the easier question correctly and the other incorrectly will have a higher total score than the subject who misses the easier question but answers the other correctly. Some reflection may be required to make this fact seem non paradoxical.

A couple of similar comments were suggested to me by F. M. Lord. One is that a subject who does well on questions 2, 3, ... should (other things being equal) receive a better score on question 1 than a subject who does poorly on questions 2, 3, ... This is presumably correct if we extend the model of this section to cover several questions and if knowledge of the answers to various questions are positively correlated.

In section 7 an extreme case of a generalization of the pair of questions model, called the scaled model, is discussed. In this case knowledge of the answer to one question implies knowledge of the answers to the preceding questions and it is relatively easy to analyze several questions simultaneously.

The reader may have noticed that the model of this section is related to that used in configural analysis (see [9]).

5. A general formulation.

The models of sections 3 and 4 suggest the following general formulation which will be applied to several other models in sections 6, 7 and 8.

Let w represent a random variable associated with a random subject. In our applications it will generally be a vector valued random variable which will characterize the knowledge of the subject and his responses. In section 4, $w = (w_1, w_2, w_3, w_4)$ could be used to describe a subject where $w_1 = 1$ if he knows the answer to the first question and 0 otherwise, $w_2 = 1$ if he knows the answer to the second question and 0 otherwise, and w_3 and w_4 are his responses to the two questions.

Let z represent the observed responses. In the above example z can be regarded as the subvector (w_3, w_4). In general z will be some function of w , i.e.,

$$z = f(w) \; . \tag{5.1}$$

We are interested in the value of the subject's knowledge since that is what is to be estimated. In the above application the value is $w_1 + w_2$ (assuming the two questions are equally important). The value can not be observed directly. It is to be estimated from the observed data represented by z . In any case the value v is also some function of w, i.e.,

$$v = g(w) \; . \tag{5.2}$$

Now we select an assigned score x to serve as the estimate of v. Since x is to depend on the observed data, we write

$$x = h(z) \; . \tag{5.3}$$

We shall select the scoring system h which minimizes the mean squared error

$$V = E\{(x-v)^2\} \; . \tag{5.4}$$

Applying the properties of conditional expectation, we note that

$$V = E\{E[(x-v)^2 | z]\} \tag{5.5}$$

and hence is minimized by selecting

(5.6) $$x = h(z) = E\{v|z\} ,$$

i.e., x is the regression of v on z. To apply this method of scoring it is necessary that enough be known about the probability distribution of w so that the above regression can be evaluated. In the models of section 3 and 4, it was possible to use the data for the population of subjects to estimate the probability distribution of w. It would suffice to know the joint distribution of v and z.

For the above method of scoring, the minimum variance is

(5.7) $$V = E\{v^2\} - E\{x^2\} .$$

For an arbitrary scoring system $x^* = h^*(z)$, the mean squared error is increased by

(5.8) $$V^* - V = E\{(x^*-x)^2\}$$

since

$$V^* = E\{(x^*-v)^2\} = E\{(x^*-x)^2\} + E\{(x-v)^2\} + 2E\{(x^*-x)(x-v)\}$$

and

$$E\{(x^*-x)(x-v)\} = E\{E[(x^*-x)(x-v)|z]\} ,$$

$$= E\{(x^*-x)\ E[(x-v)|z]\} ,$$

$$= 0 .$$

This general model and these results can be applied to the examination as a whole or to individual items or to sections of the examination. We defer some discussion till section 9 on total scores.

6. Partial Knowledge.

The simple model of section 3 does not apply in many cases. One variation introduced by Horst [1] and discussed by Gulliksen [5] and Solomon [12] allows for partial knowledge.

In Horst's model which we shall call the simple ordering of responses model, there is an order of the r possible responses. When they are arranged in order, a subject's knowledge will consist of knowing that the correct response is among the first i possible responses for some i between 1 and r. The subject selects one of these i responses at random. If λ_i denotes the portion of the population who can limit their answer to the first i responses, the proportion who select the j^{th} response is

$$p_j = \frac{\lambda_j}{j} + \frac{\lambda_{j+1}}{j+1} + \cdots + \frac{\lambda_r}{r} \qquad 1 \leq j \leq r \tag{6.1}$$

A large sample of responses will enable the investigator to estimate the p_j's directly. These will determine the appropriate order (if it is not already known) because the lower index in the appropriate order is associated with a larger p_j. Finally

$$\lambda_j = j[p_j - p_{j+1}] \qquad 1 \leq j \leq r \ . \tag{6.2}$$

In particular, we have the known result that $\lambda_1 = p_1 - p_2$, i.e., the percentage who know the answer is the difference in the observed proportions for the two most favored responses.

Let us relate this example to the general formulation of section III. For each subject let $w = (I,J)$ where I is the group from which the subject made his selection and J is his choice. The observed datum is $z = J$. The value $v = g(w)$ will ordinarily depend on the unknown I. Thus we may represent this dependence by $v = v_I$. The investigator who regards partial knowledge as worthless will let $v_1 = 1$ and $v_2 = v_3 = \cdots = v_r = 0$. For those cases where partial knowledge is regarded as valuable some other assignment can be made.

According to this model,

$$(6.3) \qquad P(w = (i,j)) = \lambda_i/i \quad \text{for } 1 \leq j \leq i , \qquad = 0 \quad \text{otherwise.}$$

To derive the optimal score $x = E(v|z)$ we first compute the conditional probability of w given $z = J = j$

$$(6.4) \qquad P(w = (i,j)|J=j) = \frac{\lambda_i/i}{\sum_{\alpha=j}^{r} \lambda_\alpha/\alpha} = \frac{\lambda_i}{ip_j} \quad \text{for } i=j, j+1,\ldots,r$$

Thus the score corresponding to $J = j$ is

$$(6.5) \qquad x_j = E(v|J=j) = \frac{1}{p_j} \sum_{i=j}^{r} \frac{\lambda_i v_i}{i}$$

and

$$(6.6) \qquad V = \sum_{i=1}^{r} \lambda_i v_i^2 - \sum_{j=1}^{r} p_j x_j^2$$

while the "standard" procedure of assigning v_j to a choice of j would lead to an increase of

(6.7) $$V^* - V = \sum_{j=1}^{r} p_j(x_j - v_j)^2$$

For the important special case where $v_1 = 1$ and $v_2 = v_3 = \cdots = v_r = 0$,

(6.5a) $$x_1 = \frac{\lambda_1}{p_1} = \frac{p_1 - p_2}{p_1} \quad ,$$

$$x_2 = x_3 = \cdots = x_r = 0 \ .$$

Here we have

(6.6a) $$V = \lambda_1 - \frac{\lambda_1^2}{p_1} = \frac{p_2(p_1 - p_2)}{p_1}$$

and

(6.7a) $$V^* - V = \frac{p_2^2}{p_1} \ .$$

The simple ordering of responses is not the only way in which partial knowledge can be represented. Other more complex models could be applied to yield what we shall call complex orderings. In general let us assume that the subject's knowledge can be represented by a set S of possible responses from which he selects one at random and among which is the correct response.[1] If $S_1, S_2, \ldots, S_r$, are possible sets one may attach values $v_1, v_2, \ldots, v_r$, to the corresponding knowledge. In some cases it is possible to use the observed frequencies of response to estimate $\lambda_1, \lambda_2, \ldots, \lambda_r$, where λ_i is defined as the proportion of subjects whose knowledge is represented by S_1. We present one example. Let $S_1 = \{a\}$, $S_2 = \{a,b\}$, $S_3 = \{a,c\}$, $S_4 = \{a,b,c,d\}$. From the observed frequencies of responses p_a, p_b, p_c, p_d, we estimate λ_1, λ_2, λ_3, and λ_4 by solving

---

[1] Still more complicated models can be constructed where the choice among the elements of S are not equally probable or even where S does not contain the correct answer.

$$(6.8) \qquad \begin{aligned} p_a &= \lambda_1 + \tfrac{1}{2}\lambda_2 + \tfrac{1}{2}\lambda_3 + \tfrac{1}{4}\lambda_4 , \\ p_b &= \tfrac{1}{2}\lambda_2 + \tfrac{1}{4}\lambda_4 , \\ p_c &= \tfrac{1}{2}\lambda_3 + \tfrac{1}{4}\lambda_4 , \\ p_d &= \tfrac{1}{4}\lambda_4 . \end{aligned}$$

While $r' \leq r$ is necessary for the λ's to be estimable, this condition is not sufficient.

If the λ's are estimable the general approach of section 5 is applicable. In those examples of complex ordering where the λ's are not consistently estimable[2], a minimax approach may be applied to the mean squared error. That is, $x = h(z)$ could be selected to minimize

$$(6.9) \qquad \max_{\theta \in \Omega} E\{(x-v)^2\}$$

where Ω is the class of all distributions of w consistent with the observed distribution of z. Note that for given θ and given $x = h(z)$

$$(6.10) \qquad E_\theta\{(x-v)^2\} = E\{[x-E_\theta(v|z)]^2\} + E_\theta\{[v-E_\theta(v|z)]^2\} .$$

---

[2] The appropriate technical term is unidentified.

The first expectation on the right hand side does not have a θ subscript because it can be computed by using the observed distribution of z. Now fix θ and let $v^*(\theta)$ represent a point in $r + 1$ dimensional space whose coordinates are $E_\theta(v|z = 1)$, $E_\theta(v|z = 2)$, $\ldots$, $E_\theta(v|z = r)$, $[E_\theta\{[v - E_\theta(v|z)]^2\}]^{1/2}$. Let x^* be the point whose coordinates are $x_1 = h(1)$, $x_2 = h(2)$, $\ldots$, $x_r = h(r)$, 0. Then $E_\theta\{(x-v)^2\}$ is the squared distance between x^* and $v^*(\theta)$. The minimax estimate has a simple geometric interpretation. It corresponds to the point on a given r dimensional plane in $r + 1$ dimensional space for which the maximum distance to the points of a specified set is as small as possible. The specified set is $\{v^*(\theta) : \theta \in \Omega\}$. We shall not discuss complex orderings further.

7. Questions which can be scaled.

In section 4 we discussed a model where the results on two questions were combined. One object was to capitalize on the correlation between the knowledge of the answers to several questions. The analysis for two questions was more difficult than for a single question. We indicated two objections to the extension to many questions. These were the difficulty of analysis and the large sample size required to obtain good estimates of the necessary parameters.

There is a special extreme case for which these objections do not apply. This is the case where the questions test knowledge which can be scaled. More precisely, suppose that there are s questions such that any subject who knows the answer to the i^{th} question knows the answer to all the preceding ones. If he knows the answer to only the first i, he answers them correctly but guesses at random among the r choices on each of the subsequent questions. Let λ_i be the proportion of subjects in this category for $i = 1, 2, \dots, s$ and let λ_0 be the proportion who don't know the answer to any of the questions. Let

$$w = (I, J, K_1, K_2, \dots, K_L) \tag{7.1}$$

where $I \geq 0$ indicates the number of questions whose answers the subject knows, $J \geq I$ indicates the number of correct answers before a wrong one appears and $K_1, K_2, \dots, K_L$ are the other questions, if any, which are answered correctly. The subject's response may be represented by

$$z = (J, K_1, K_2, \dots, K_L) \quad . \tag{7.2}$$

The value associated with the subjects knowledge can be represented by $v = v_I$. One may use $v = I$ although other measures may be appropriate on occasion.

We have

$$P\{w = (i,j,k_1,k_2, \dots, k_\ell)\} = \lambda_i \left(\frac{1}{r}\right)^{j+\ell-i} \left(\frac{r-1}{r}\right)^{s-j-\ell} , \tag{7.3}$$

$$P\{z = (j,k_1,k_2, \dots, k_\ell)\} = \sum_{i=0}^{j} \lambda_i \, r^{-(s-i)} (r-1)^{s-j-\ell} , \tag{7.4}$$

$$P\{w = (i,j,k_1, \dots) \mid z = (j,k_1, \dots)\} = \lambda_i \, r^i \Big/ \sum_{i=0}^{j} \lambda_i \, r^i \qquad \text{for } 0 \leq i \leq j . \tag{7.5}$$

Then, given $z = (j,k_1,k_2, \dots, k_\ell)$, the appropriate score is

$$x_j = \sum_{i=0}^{j} v_i \, \lambda_i \, r^i \Big/ \sum_{i=0}^{j} \lambda_i \, r^i \tag{7.6}$$

where λ_i may be estimated using the fact that

(7.7) $$p_j = P\{J = j\} = \frac{r-1}{r^{j+1}} \sum_{i=0}^{j} \lambda_i r^i .$$

and hence

(7.8) $$\lambda_j = (r-1)^{-1} \{rp_j - p_{j-1}\}$$

where the p_j may be directly estimated from the observed frequencies of the random variable J. As may have been expected, $K_1, K_2, \cdots, K_L$ are not involved in the estimation of λ and do not influence x.

The mean squared error corresponding to the optimal procedure is given by

(7.9) $$V = \sum_{i=0}^{r} \lambda_i v_i^2 - \sum_{j=0}^{r} p_j x_j^2 .$$

8. A model involving reexamination.

An example which was brought to the author's attention consisted of a Spanish Language examination which was given to a group of subjects before and after special training. Each question offered three choices and it seemed clear from the nature of the examination that the simple ordering partial knowledge model was appropriate. On the other hand one would expect a definite relationship between the knowledge before and after the training. That is, one would expect that on the whole, students would know at least as much for the second examination as they would for the first. Then it would make sense to consider a question before training and one after training as a pair of questions to which the models of scaling and simple ordering could be applied.

However data were unavailable to substantiate the expectation that subjects did not forget between the examinations which were to be given six months apart. Therefore it was decided to drop the scaling assumption out of fear that if it did not hold, the results would be questionable. The following model was applied.

Let $1, 2, \ldots, r$ be the natural order of choice among the possible responses. Let $S_1 = \{1\}$, $S_2 = \{1,2\}$, $S_3 = \{1,2,3\}, \ldots$ represent the possible knowledge of a subject responding to the question. We represent the combined knowledge and response of a subject to the question on both examinations by

$$w = (I_1, I_2, J_1, J_2) \tag{8.1}$$

where the subject knows that the answer is in S_{I_1} on the first examination, knows that the answer is in S_{I_2} on the second examination, and responds J_1 on the first and J_2 on the second examination. Let v_i be the value associated with S_i. Thus the subject deserves a value $v_{(1)} = v_{I_1}$ on the first examination a value $v_{(2)} = v_{I_2}$ on the second. In the above mentioned Spanish examination, the investigator was satisfied only with complete knowledge, i.e. $v_1 = 1$, $v_2 = v_3 = 0$. The responses are denoted by $z = (J_1, J_2)$.

Let

$$\lambda_{i_1, i_2} = P\{I_1 = i_1, I_2 = i_2\} . \tag{8.2}$$

Then

$$(8.3)\quad P\{w = (i_1,i_2,j_1,j_2)\} = \frac{\lambda_{i_1,i_2}}{i_1 i_2} \quad \text{for} \quad 1 \leq j_1 \leq i_1,\ 1 \leq j_2 \leq i_2 ,$$

and

$$p_{j_1,j_2} = P\{z = (j_1,j_2)\} = \sum_{\substack{j_1 \leq i_1 \leq r \\ j_2 \leq i_2 \leq r}} \lambda_{i_1,i_2} (i_1\, i_2)^{-1}$$

The appropriate scores to be assigned to a subject for the two tests are given by

$$(8.4)\qquad x_1 = E\{v_{I_1} | z = (j_1,j_2)\}$$

and

$$(8.5)\qquad x_2 = E\{v_{I_2} | z = (j_1,j_2)\} .$$

If $z = (j_1,j_2)$,

$$x_1 = (p_{j_1,j_2})^{-1} \left\{ \sum_{\substack{j_1 \leq i_1 \leq r \\ j_2 \leq i_2 \leq r}} v_{i_1} \lambda_{i_1,i_2} (i_1 i_2)^{-1} \right\} .$$

However

$$\sum_{i_2=j_2}^{r} \lambda_{i_1,i_2} (i_1 i_2)^{-1} = p_{i_1,j_2} - p_{i_1+1,j_2}$$

where a p with $r+1$ for a subscript is assumed to be zero. It follows that if $z = (j_1,j_2)$

(8.6) $$x_1 = (p_{j_1,j_2})^{-1}\left\{\sum_{i_1=j_1}^{r} v_{i_1}[p_{i_1,j_2} - p_{i_1+1,j_2}]\right\}$$

and

(8.7) $$x_2 = (p_{j_1,j_2})^{-1}\left\{\sum_{i_2=j_2}^{r} v_{i_2}[p_{j_1,i_2} - p_{j_1,i_2+1}]\right\} .$$

It may be noted that $(i_1\ i_2)^{-1}\lambda_{i_1,i_2}$ can be computed by taking second differences. Then

(8.8) $$\lambda_{i_1,i_2} = (i_1 i_2)[p_{i_1,i_2} - p_{i_1+1,i_2} - p_{i_1,i_2+1} + p_{i_1+1,i_2+1}] .$$

We illustrate with an artificial example with $r = 3$, $v_1 = 1$, $v_2 = .2$, $v_3 = 0$, where the frequency of observations (J_1, J_2) is given by the following table which also includes the marginal distribution

$p_{j_1,\ j_2}$

$j_1 \backslash j_2$	1	2	3	
1	.340	.130	.035	.505
2	.180	.100	.025	.205
3	.100	.070	.020	.190
	.620	.300	.080	

The tables for $p_{i_1,j_2} - p_{i_1+1,j_2}$ and $p_{j_1,i_2} - p_{j_1,i_2+1}$ are given below.

$p_{i_1,j_2} - p_{i_1+1,j_2}$

$i_2 \backslash j_2$	1	2	3
1	.160	.030	.010
2	.080	.030	.005
3	.100	.070	.020

$p_{j_1,i_2} - p_{j_1,i_2+1}$

$i_2 \backslash j_2$	1	2	3
1	.210	.095	.035
2	.080	.075	.025
3	.030	.050	.020

We compute x_1 and x_2 for each possible value (j_1, j_2) of z.

x_1

j_1 \ j_2	1	2	3
1	.518	.277	.314
2	.089	.060	.040
3	.000	.000	.000

x_2

j_1 \ j_2	1	2	3
1	.674	.146	.000
2	.528	.150	.000
3	.400	.142	.000

The table for λ_{i_1,i_2} and the associated marginal probabilities $\lambda_{i_1 \cdot}$ and $\lambda_{\cdot i_2}$ is also presented below

λ_{i_1,i_2}

i_1 \ i_2	1	2	3	
1	.130	.040	.030	.200
2	.100	.100	.030	.230
3	.090	.300	.180	.570
	.320	.440	.240	

In specific applications it may pay to compute the λ's merely to determine whether the model is consistent with the data. If some of the λ's are negative one may ask whether the model fails to apply or whether sampling variation was responsible for the discrepancy.

We may compute

$$V_1 = E\{v_{(1)}^2\} - E\{x_1^2\} = .1029 \tag{8.9}$$

and

$$V_2 = E\{v_{(2)}^2\} - E\{x_2^2\} = .1107 \tag{8.10}$$

In this example the mean squared error derived from using the naive scores x_1^* and x_2^* which assign value v_j to response j are given by

$$V_1^* = E\{(x_1^* - v_1)^2\} = .2712 \tag{8.11}$$

and

$$V_2^* = E\{(x_2^* - v_2)^2\} = .2240 \ . \tag{8.12}$$

Finally, if the relationship between the two questions is ignored and the simple ordering partial knowledge model is applied to each we would have scores given by $x_1^{**} = .442$, $.075$, and 0 for $J_1 = 1,2,$ and 3 respectively and $x_2^* = .587$, $.147$ and 0 for $J_2 = 1, 2,$ and 3. The mean squared error for this procedure is given by

$$V_1^{**} = .1090$$

and

$$V_2^{**} = .1174 \ .$$

In this example, combining the two questions does not lead to much improvement over considering the questions separately.

The special case where $v_1 = 1$, $v_2 = v_3 = \cdots = v_r = 0$ gives rise to much simpler formulae for the scores. Here, if $z = (j_1, j_2)$, we have

$$x_1 = (p_{1,j_2})^{-1} (p_{1,j_2} - p_{2,j_2}), \qquad \text{if } j_1 = 1, \tag{8.6a}$$

$$x_1 = 0 \qquad \text{if } j_1 \neq 1,$$

$$x_2 = (p_{j_1,1})^{-1} (p_{j_1,1} - p_{j_1,2}), \qquad \text{if } j_2 = 1, \tag{8.7a}$$

$$x_2 = 0 \qquad \text{if } j_2 \neq 1,$$

$$V_1 = \lambda_{1\cdot} - \sum_{j_2=1}^{r} (p_{1,j_2})^{-1} (p_{1,j_2} - p_{2,j_2})^2 \quad , \tag{8.9a}$$

$$V_2 = \lambda_{\cdot 1} - \sum_{j_1=1}^{r} (p_{j_1,1})^{-1} (p_{j_1,1} - p_{j_1,2})^2 \quad . \tag{8.10a}$$

9. Composite scores.

Consider an examination consisting of a considerable number of individual questions. How should a subject be assigned a composite score for the examination as a whole ? At several points in this paper it may have seemed to be implicitly implied, that simply totaling the scores for the individual questions or items in the examination is the proper and natural procedure. Indeed, this procedure is far from unreasonable and it will be evaluated incidentally later in this section. However it is important to note that the method of totaling, innocent though it seems, is ordinarily non-optimal. This non-optimality arises principally from the possible interdependence of different questions. Thus knowledge of the answer of one question is ordinarily correlated with knowledge of the answer to another. Another issue, which arises and is briefly discussed at the end of this section, is the possibility that composite scores are desired, not so much to evaluate a subject's knowledge, as, to discriminate among subjects of different ability.

Let us digress briefly to point out that, if it is desired to evaluate a subject's knowledge, the general formulation of section 5 applies to the entire examination considered as a unit. However, unless the examination consists of few questions or of a few sections to each of which the scale model of section 7 is applicable, practical difficulties would make it unfeasible to apply the method of section 5. In fact these difficulties are exactly those mentioned in section 4 where it was indicated that it would be unfeasible to extend the two question model very far.

How should we proceed if we discard the idea of applying the general formulation to the entire examination? One approach that has been applied consists of adopting some strongly parametrized assumptions and then applying standard techniques such as maximum-likelihood to the data. In such cases the fundamental approach of this paper is essentially irrelevant. Maximum-likelihood estimation has been applied with forms of the probit and logit model by Lord [8] and Birnbaum [1,2,3]. Although they were not mainly interested in multiple choice questionnaires, these were considered. On the other hand the models they applied avoided an important aspect of the possibility of interdependence of different questions.

A second approach which we propose here is to let the composite score be the sum of individual scores which are assigned so as to minimize the mean squared deviation of the sum from the overall value of the subject's knowledge.

To illustrate let us apply this approach to the case of an examination consisting of k questions, for each of which the simplest model with r choices applies. As before p_i and λ_i represent respectively the proportions of the population of subjects who obtain the correct response and who know the answer to the i^{th} question. Also

$$\lambda_{ij} = \lambda_i \lambda_j + \theta_{ij} \tag{9.1}$$

represents the proportion who know the answer to both the i^{th} and j^{th} questions and can be estimated by observing p_i, p_j, and p_{ij}, the proportion who answer both questions correctly. In fact

$$p_{ij} = p_i p_j + \theta_{ij} [1 - \frac{1}{r}]^2 \tag{9.2}$$

if we assume that responses to questions for which the subject doesn't know the answer are random and independent of one another. Let us also assume that the value of a subject's knowledge can be expressed by

$$v = \sum_{i=1}^{k} v_i \tag{9.3}$$

where $v_i = a_i$ if the subject knows the answer to the i^{th} question and 0 otherwise.

For the sake of convenience, we may express the subject's total or composite score by

$$x^* = \delta_o + \sum_{i=1}^{k} (x_i+u_i) = \delta_o + \sum_{i=1}^{k} x_i^* \tag{9.4}$$

where $x_i^* = x_i+u_i$, $u_i = \delta_i$ if the i^{th} answer is correct and 0 otherwise, and x_i is the appropriate score for the single question model. That is, $x_i = a_i\lambda_i/p_i$ for a correct answer and 0 otherwise. Then the mean squared deviation of the composite score from the value of the subject's knowledge is given by

$$\begin{aligned} V^* &= E\left\{\delta_o + \Sigma(x_i^*-v_i)\right\}^2 \\ &= (\delta_o + \sum_i \delta_i p_i)^2 + E\left\{\sum_i (x_i-v_i)+(u_i-\delta_i p_i)\right\}^2 . \end{aligned}$$

Now

$$\sum_i E[(x_i - v_i) + (u_i - \delta p_i)]^2 = \sum_i E[x_i - v_i]^2 + \sum_i E[u_i - \delta_i p_i]^2$$

$$= V + \sum_i p_i(1-p_i)\delta_i^2$$

where

$$V = \sum_i a_i^2 \lambda_i \left(1 - \frac{\lambda_i}{p_i}\right) \tag{9.5}$$

represents the sum of the mean squared deviations for optimal responses to individual questions. Finally, for $i \neq j$,

$$E[(x_i - v_i + u_i - \delta_i p_i)(x_j - v_j + u_j - \delta_j p_j)] = E[(x_i^* - v_i)(x_j^* - v_j)] - \delta_i \delta_j \, p_i p_j$$

$$= \theta_{ij}\left[\frac{a_i}{rp_i} - \delta_i\left(1 - \frac{1}{r}\right)\right]\left[\frac{a_j}{rp_i} - \delta_j\left(1 - \frac{1}{r}\right)\right]$$

and we have

$$V^* = V + \left(\delta_o + \sum_i \delta_i p_i\right)^2 + \sum_i p_i(1-p_i)\,\delta_i^2 + \sum_{i \neq j} \theta_{ij}\left[\frac{a_i}{rp_i} - \delta_i\left(1 - \frac{1}{r}\right)\right]\left[\frac{a_j}{rp_i} - \delta_j\left(1 - \frac{1}{r}\right)\right] \tag{9,6}$$

It follows that in the case of independence, $\theta_{ij} = 0$ for $i \neq j$, the optimal score would be $x = \Sigma x_i$. In fact the mean squared error for x is easily obtained, even when $\theta_{ij} \neq 0$, by setting all the δ_i equal to zero in equation (9.6)

Note that in the general case δ_o should be selected to make the mean error zero. With this adjustment, it is desirable to minimize

$$\sum_i p_i(1-p_i)\,\delta_i^2 + \sum_{i \neq j} \theta_{ij}\left[\frac{a_i}{rp_i} - \delta_i\left(1 - \frac{1}{r}\right)\right]\left[\frac{a_j}{rp_j} - \delta_j\left(1 - \frac{1}{r}\right)\right] \; .$$

whose partial derivative with respect to δ_{i_o} is given by

$$2\,\delta_{i_o} p_{i_o}(1 - p_{i_o}) - 2\left(1 - \frac{1}{r}\right) \sum_{j \neq i_o} \theta_{i_o j}\left[\frac{a_j}{rp_j} - \delta_j\left(1 - \frac{1}{r}\right)\right] \; .$$

Hence if the θ_{ij} are small and positve, as is often assumed to be the case, the optimal δ_i would tend to be positive.

It should be emphasized that in the procedure evaluated above we have severely limited our choice of scoring methods and we have no simple way of evaluating how serious a loss this limitation imposes.

Thus far we have assumed that the main objective of the examination score is to evaluate the subject's knowledge. These are occasions when this is not the primary objective. In many cases where a single composite score is desired the objective is to discriminate among students of different ability. Then, quite a different approach to scoring may be appropriate. For example, we propose here to use as our discriminant function a composite score of the form

$$D = \sum_i \alpha_i x_i \tag{9.7}$$

where the x_i is the score which would be appropriate for estimating v_i, the value the i^{th} part of an examination consisting of k independent parts. In fact let us select the α_i so as to minimize

$$V_D = \sum_i \alpha_i^2 \, E(x_i - v_i)^2 \tag{9.8}$$

subject to the condition that

$$\mu_D = \Sigma \alpha_i \, E(v_i) \tag{9.9}$$

be a fixed specified value. This criterion leads to the use of

$$\alpha_i = E(v_i)/E(x_i - v_i)^2 \quad . \tag{9.10}$$

This proposal is subject to criticism. An argument could be made for it if one were to assume that the difference between two subjects were expressed by a factor θ so that the ratios of their v_i's, v_{i1}/v_{i2} is approximately θ.

This last assumption is questionable. For individual questions the ratio v_{i1}/v_{i2} can vary enormously. Even in considering groups of questions v_{i1}/v_{i2} will tend to be one for a group of easy questions and may be quite different for a group of difficult questions. The assumption of independence is also questionable in practice.

Aside from doubts about the assumptions, one should note that for easy questions, $E(x_i-v_i)^2$ may be small compared to $E(v_i)$. Then a subject who missed an easy question would suffer severely. While there is justification for this, one should consider tempering the "punishment" in case the question was missed for reasons outside the model such as a slip of the pen or an oversight.

Finally note that the equations of this section apply even in those cases where stratification of the population preceded the application of scoring.

10. Miscellaneous Remarks.

The method proposed in this paper involves a combination of two ideas. First, the performance of the population can be used to estimate the distribution of knowledge or ability among the population. Second, knowing the distribution of knowledge among the population one may apply minimum mean squared error estimation to estimate the subject's knowledge.

This type of combination appeared in a problem given on matriculation examinations at Columbia University and Stanford University. This problem originated, I believe, with Howard Raiffa, as a means of exposing the foundation of Robbins' Theory of Empirical Bayes Procedures. It evolved to the following form "A tubetester always reads good for a good tube and reads good for 25% of the bad tubes. A million tubes are tested. The cost of rejecting a good tube is equal to that of accepting a bad tube. What procedure is called for if 240,000 of the tubes receive good readings?" This problem seemed rather difficult and exotic to many students.

The possibility of using the performance of the population to help score individuals arose in a method suggested by Hamilton [6] to correct for guessing and in a predecessor of this method due to Calandra [4]. As we indicated in section 3, this use was subjected to some criticism.

The idea of applying minimum mean squared error, and consequently regression theory, was previously introduced by Sitgreaves [11] in a different model, which involved a combination of the probit type of model used by Lord [8] and that of questions which were scaled.

When enough of a parametric structure is put into the model it becomes feasible to attack the problem of design of questionnaires. The author has done some preliminary work for a model where the probability that an individual know the answer to the i^{th} question is

$$\lambda_i = 1-e^{-\theta t_i}$$

with t_i a measure of the "ease" of the question and θ a measure of the ability of the subject. Similar work was done for multiple choice questionnaires incidentally in the attack on the design of non-multiple choice questions by Lord [8], Birnbaum [3]. The general problem of design is one of the principal subjects of Studies in Item Analysis and Prediction," edited by Solomon [13].

The validity of a test plays an important role in most studies of item analysis. In this paper we have ignored this question and there is no essental need to introduce it. The fundamental issue here concerns how to estimate what the test does measure. The question of whether the test measures what we would like it to can be treated separately. In fact if a study were to show how a criterion C is related to knowledge of the answers to items on the test, this study could presumably be used to derive good values v_i for portions of the test.

BIBLIOGRAPHY

[1] A. Birnbaum, "Efficient design and use of tests of a mental ability for various decision-making problems," Technical Report, USAF School of Aviation Medicine Project No 7755-23, Teechers College, Columbia University, Jan. (1957) pp. 1-25.

[2] A. Birnbaum, "On the estimation of mental ability," Series Report No. 15. USAF School of Aviation Medicine Project No. 7755-23, Teachers College, Columbia University, January (1957) pp. 1-31.

[3] A. Birnbaum, "Further considerations of efficiency in tests of a mental ability," Technical Report No. 17, USAF School of Aviation Medicine Project No 7755-23, Teachers College, Columbia University, pp. 1-39.

[4] A. Calandra, "Scoring formulas and probability considerations," Psychometrika, Vol. 6, (1941) pp. 1-9.

[5] H. Gulliksen, Theory of Mental Tests, New York, John Wiley and Sons, Inc., (1950) pp. 1-486.

[6] C. H. Hamilton, "Bias and error in multiple choice tests," Psychometrika, Vol. 15, (1950) pp. 151-168.

[7] P. Horst, "The difficulty of a multiple choice test item," Journ. of Educational Psych., Vol. 24, (1933) pp. 229-232.

[8] F. M. Lord, "An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability, Psychometrika, Vol. 18, No. 1, (1953) pp. 57-76.

[9] A. Lubin and H. G. Osburn, "A theory of pattern analysis for the prediction of a quantitative criterion," Psychometrika, Vol. 22, No. 1, (1957) pp. 63-73.

[10] S. B. Lyerly, "A note on correcting for chance success in objective tests," Psychometrika, Vol. 16, (1951) pp. 21-30.

[11] R. Sitgreaves, "Optimal test design in a special testing situation," Studies in Item Analysis and Prediction, Edited by H. Solomon, Stanford University Press (1961) pp. 29-45.

[12] H. Solomon, "Item analysis and classification techniques," Third Berkeley Symposium on Math. Stat. and Prob., Vol. 5, Edited by J. Neyman, University of California Press (1955) pp. 169-184.

[13] H. Solomon, Studies in Item Analysis and Prediction, Stanford University Press (1961) pp. 1-324.

THE USE AND INTERPRETATION OF PRINCIPAL COMPONENT ANALYSIS IN APPLIED RESEARCH

by

C. Radhakrishna Rao

1. Introduction.

During the last few years, I had the opportunity of surveying some of the multivariate statistical techniques and evaluating their usefulness in practical research work (Rao, 1960, 1961). I undertook to do this mainly because the superiority of the multivariate methods over the simpler univariate analysis has been questioned by some statisticians. It was said that an examination of individual measurements by simpler univariate methods is an essential step which can be supplemented by more elaborate multivariate analysis when this appears feasible and necessary. It was also thought that multivariate methods have only provided new outlets for mathematical theory without materially assisting scientific research.

It is, indeed, necessary to criticise uncritical and inappropriate applications, which are perhaps natural when new techniques are put forward. But a criticism of multivariate methods as such is not justifiable. I have considered a number of practical examples specially in one of the papers (Rao, 1961) to demonstrate the use of multivariate analysis and to emphasize the need for exploring the potenialities of the existing techniques and for further research.

The purpose of the present paper is to examine, in some detail, the role of principal component analysis in applied research. When a large number of measurements are available it is natural to enquire whether they could be replaced by a fewer number of the measurements or of their functions, *without loss of much information*, for convenience in the analysis and in the interpretation of data. Principal components, which are linear functions of the measurements, are suggested for this purpose. It is, therefore, relevant to examine in what sense principal components provide a reduction of the data without much loss of *information we are seeking from the data*.

The following are some typical statements found in the literature on applications of principal component analysis: 'The transformed variables (the first few principal components) are useful in connection with preliminary investigation of a large number of samples of species from different localities'. 'For an orientatory type of investigation, it would probably suffice with the variable provided by the first principal component (out of four) for this represents 3/4 of the total variation'. 'At first principal components are most likely to be regarded by the non-mathematicians as a highly arbitrary set of manipulations. Such a reaction sould be dismissed, as soon as the geometrical meaning is considered: principal component analysis merely leads to new angles of viewing data, analysis best suited to disclose the nature of size and shape variation'. In no case are the statements made substantiated by statistical analysis to show that the information neglected does not lead to misleading conclusions. Often, the principal component analysis is first undertaken, without any clear objective, pooling all the available data and then an attempt is made to interpret the derived results.

The discussion in the present paper has the following aims: (a) To provide various interpretations of principal components. (b) To examine the situations where the principal component analysis can be undertaken with a definite purpose or in an exploratory way in the earlier stages of investigation of a research problem. (c) To indicate the difference between the principal component analysis and the factor analysis. (d) To generalize the principal component analysis in a number of directions useful in applied research.

2. Eigen values and vectors of matrices.

For a theoretical development of the principal component analysis and its interpretation it is necessary to use some results on the canonical reduction of matrices, which are summarized in this section for use in the later sections.

(i) Eigen values and vectors of a matrix. Let $\underline{\Sigma}$ be a non-negative (i.e., positive definite or positive semi-definite) matrix of order p × p. The roots of the determinental equation

$$|\underline{\Sigma} - \lambda \underline{I}| = 0 \tag{2.1}$$

are called the eigen values of $\underline{\Sigma}$. The equation (2.1) has p real non-negative roots, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. Corresponding to each root λ_i of (2.1), there exists a vector $\underline{P}_i$, such that

$$\underline{\Sigma}\, \underline{P}_i = \lambda_i\, \underline{P}_i \,, \tag{2.2}$$

which is called an eigen vector. Then the following results hold:

(a) $\underline{P}_1,\ldots,\underline{P}_p$ can be chosen to be orthonormal whether $\lambda_1,\ldots,\lambda_p$ are distinct or not.

(2.3) (b) $\underline{\Sigma} = \lambda_1 \underline{P}_1 \underline{P}_1' + \cdots + \lambda_p \underline{P}_p \underline{P}_p'$.

$$\underline{I} = \underline{P}_1 \underline{P}_1' + \cdots + \underline{P}_p \underline{P}_p' \quad .$$

(c) Let $\underline{L}_1, \ldots, \underline{L}_q$ be any set of orthonormal vectors. Then

$$\sum_{i=1}^{q} \underline{L}_i' \, \underline{\Sigma} \, \underline{L}_i \leq \sum_{1}^{q} \underline{P}_i' \, \underline{\Sigma} \, \underline{P}_i = \lambda_1 + \cdots + \lambda_q \tag{2.4}$$

$$q = 1, \ldots, p \; .$$

(d) Let $\underline{B}$ be any matrix of rank q. Then

$$\min_{\underline{B}} \|\underline{\Sigma} - \underline{B}\| = (\lambda_{q+1}^2 + \cdots + \lambda_p^2)^{1/2} \tag{2.5}$$

and the minimum is attained when

$$\underline{B} = \lambda_1 \underline{P}_1 \underline{P}_1' + \cdots + \lambda_q \underline{P}_q \underline{P}_q' \; . \tag{2.6}$$

In (2.5), the symbol $\|\underline{A}\|$ denotes the Euclidean norm, which is the square root of the sum of squares of the elements of $\underline{A}$. Or in other words, the *choice of $\underline{B}$ as in (2.6) is the best fitting matrix of given rank q to $\underline{\Sigma}$.*

(ii) Eigen values and vectors associated with a pair of matrices.

Let $\underline{\Sigma}$ and $\underline{\Gamma}$ be two symmetric matrices of order $(p \times p)$ such that $\underline{\Gamma}$ is positive definite. Consider the determinental equation

$$|\underline{\Sigma} - \lambda \underline{\Gamma}| = 0 \; . \tag{2.7}$$

The equation (2.7) has p roots, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ which may be called eigen values of $\underline{\Sigma}$ with respect to $\underline{\Gamma}$. Corresponding to each root λ_i there is a vector $\underline{P}_i$ such that

$$\underline{\Sigma}\,\underline{P}_i = \lambda_i\,\underline{\Gamma}\,\underline{P}_i\ ,$$

which is called an eigen vector. Then the following results hold:

(a) $\underline{P}_1,\ldots,\underline{P}_p$ can be chosen such that

$$\underline{P}_i'\,\underline{\Gamma}\,\underline{P}_i = 1\ ,\ \underline{P}_i'\,\underline{\Gamma}\,\underline{P}_j = 0\ ,\ i \neq j \qquad i,j=1,\ldots,p \tag{2.8}$$

$$\text{(b)}\quad \underline{\Sigma} = \lambda_1\,\underline{P}_1\,\underline{P}_1' + \cdots + \lambda_p\,\underline{P}_p\,\underline{P}_p' \qquad \underline{\Gamma} = \underline{P}_1\,\underline{P}_1' + \cdots + \underline{P}_p\,\underline{P}_p' \tag{2.9}$$

(c) Let $\underline{L}_1,\ldots,\underline{L}_q$ be any set of vectors satisfying the same conditions as $\underline{P}_i$ in (2.8). Then

$$\sum_{i=1}^{q}\underline{L}_i'\,\underline{\Sigma}\,\underline{L}_i \leq \sum_{i=1}^{q}\underline{P}_i'\,\underline{\Sigma}\,\underline{P}_i = \lambda_1 + \cdots + \lambda_q \qquad q=1,\ldots,p\ . \tag{2.10}$$

(iii) The Eigen values and vectors under restrictions. Consider a symmetric $p \times p$ matrix $\underline{\Sigma}$ and a $p \times k$ matrix $\underline{C}$ of rank k. Let $\underline{L}_1,\ldots,\underline{L}_q$ be p dimensional vectors satisfying the conditions

$$\begin{cases} \text{(a)}\ \ \underline{L}_i'\,\underline{L}_i = 1\ ,\ \underline{L}_i'\,\underline{L}_j = 0 \ \text{ for } i \neq j \\ \text{(b)}\ \ \underline{L}_i'\,\underline{C} = 0\ \ ,\ i = 1,\ldots,q\ . \end{cases} \tag{2.11}$$

Then the maximum of

$$\underline{L}_1'\,\underline{\Sigma}\,\underline{L}_1 + \cdots + \underline{L}_q'\,\underline{\Sigma}\,\underline{L}_q \tag{2.12}$$

is attained when $\underline{L}_i = \underline{R}_i$, the i^{th} eigen vector of the matrix $(\underline{I}-\underline{C}(\underline{C}'\ \underline{C})^{-1}\underline{C}')\underline{\Sigma}$. The maximum value of (2.12) is then $\nu_1 + \dots + \nu_q$ where ν_i is the i^{th} eigen value of the matrix $(\underline{I}-\underline{C}(\underline{C}'\ \underline{C})^{-1}\underline{C}')\underline{\Sigma}$. It is easy to see that if ν_i, $\underline{R}_i$ are the eigen values and vectors of $(\underline{I}-\underline{C}(\underline{C}'\ \underline{C})^{-1}\underline{C}')\underline{\Sigma}$ and μ_i, $\underline{M}_i$ are the eigen values and vectors of the symmetric matrix $\underline{\Sigma}^{1/2}(\underline{I}-\underline{C}(\underline{C}'\ \underline{C})^{-1}\underline{C})\underline{\Sigma}^{1/2}$, then $\lambda_i = \mu_i$ and $R_i = \gamma_i\ \underline{\Sigma}^{-1/2}\underline{M}_i$, where $\gamma_i = \underline{M}'_i\ \underline{\Sigma}^{-1}\ \underline{M}_i$. Thus the eigen vectors R_i are obtained from the eigen vectors of a symmetric matrix.

(iv) Relation between the eigen values and vectors of the matrices AA' and A'A. Let $\underline{A}$ be a matrix of order $n \times p$, and rank r. Then $\underline{A}\ \underline{A}'$ and $\underline{A}'\underline{A}$ are both symmetric and have the same non-zero eigen values $\lambda_1, \dots, \lambda_r$. The multiplicity of the zero root is $n-r$ for $\underline{A}\ \underline{A}'$ and $p-r$ for $\underline{A}'\underline{A}$. Further let P_i be the eigen vector of $\underline{A}\ \underline{A}'$ and $\underline{Q}_i$ be the eigen vector of $\underline{A}'\underline{A}$ corresponding to the same non-zero eigen value λ_i. Then $\underline{Q}_i = \lambda_i^{-1/2}\underline{A}'\underline{P}_i$ and $\underline{P}_i = \lambda_i^{-1/2}\underline{A}\ \underline{Q}_i$, so that $\underline{P}_i$ can be obtained from $\underline{Q}_i$ and vice versa.

3. Fitting a subspace to a set of points in a higher dimensional space.

Principal component analysis involving the application of eigen values and vectors of a matrix was first encountered by Karl Pearson (1901) and Frisch (1929) in the problem of fitting a line, a plane or in general a subspace to a scatter of points in a higher dimensional space. Let

$$\underset{\sim}{X} = \begin{pmatrix} X_{11} & \dots & X_{1n} \\ \cdot & \cdot\ \cdot\ \cdot & \cdot \\ X_{p1} & \dots & X_{pn} \end{pmatrix} \tag{3.1}$$

represent n points (each column vector representing a point) in a p dimensional space. Any typical point is represented by $\underline{X}$. The i^{th} point is $\underline{X}_i$ and the center of gravity of the points is $\underline{\bar{X}}$, with the s^{th} coordinate

$$(3.2) \qquad \bar{X}_s = \sum_{j=1}^{n} X_{sj} \div n \, .$$

The points $\underline{X}_i$ as measured from the center of gravity are

$$(3.3) \qquad \underset{\sim}{X}_d = \begin{pmatrix} X_{11}-\bar{X}_1 & \cdots & X_{1n}-\bar{X}_1 \\ \cdot & \cdots & \cdot \\ X_{p1}-\bar{X}_p & \cdots & X_{pn}-\bar{X}_p \end{pmatrix} .$$

Let us define

$$(3.4) \qquad \underline{\Sigma} = \underset{\sim}{X}_d \, \underset{\sim}{X}_d' = (\sigma_{ij})$$

as the total dispersion matrix (or scatter) of the points. In (3.4) the element σ_{ij} is computed by the formula

$$\sigma_{ij} = \sum_{r=1}^{n} (X_{ir}-\bar{X}_i)(X_{jr}-\bar{X}_j) = \sum_{r=1}^{n} X_{ir}X_{jr} - n\,\bar{X}_i\bar{X}_j \, .$$

Now a q dimensional subspace is specified by a point (origin) and q orthogonal axes (each axis specified by its direction cosines) passing through it. Pearson defined a best fitting subspace as that for which the sum of squares of the perpendiculars from the points to the subspace is a minimum. It is easily shown that such a subspace passes through the center of gravity of the points. Further, the sum of squares of the perpendiculars from the points to the subspace defined by the center of gravity and q orthonormal vectors, $\underline{L}_1, \ldots, \underline{L}_q$, is

$$(3.5) \qquad \sum_{i=1}^{p}\sigma_{ii} - \sum_{i=1}^{q}\underline{L}_i'\,\underline{\Sigma}\,\underline{L}_i \quad .$$

Minimizing (3.5) is the same as maximizing $\sum \underline{L}_i'\,\underline{\Sigma}\,\underline{L}_i$ and an application of the result (2.4) shows that the maximum is attained when

$$(3.6) \qquad \underline{L}_i = \underline{P}_i \quad , \quad i=1,\ldots,q$$

where $\underline{P}_1,\ldots,\underline{P}_q$ are the first q eigen vectors of the matrix $\underline{\Sigma}$.

What we need in practice is an actual representation of points in the best fitting lower dimensional subspace. This is simply done by computing for each $\underline{X}$, the q coordinates

$$(3.7) \qquad \underline{P}_1'\underline{X},\ldots,\underline{P}_q'\underline{X} \quad .$$

Thus the points in the projected space referred to q orthogonal axes are

$$(3.8) \qquad \underset{\sim}{Y} = \begin{pmatrix} \underline{P}_1'\underline{X}_1 & \cdots & \underline{P}_1'\underline{X}_n \\ . & \cdots & . \\ \underline{P}_q'\underline{X}_1 & \cdots & \underline{P}_q'\underline{X}_n \end{pmatrix} .$$

The first row of $\underset{\sim}{Y}$ gives the best one dimensional representation, the first two rows of $\underset{\sim}{Y}$ give the best two dimensional representation, and so on.

In practice, it is also necessary to know how good is the representation of points in a lower dimensional space. A suitable criterion for this purpose is the sum of squares of the distances between the original and the projected points. For the best q space, the criterion has the value

$$(3.9) \qquad \lambda_{q+1} + \cdots + \lambda_p$$

using the result (2.4). Instead of the absolute value (3.9), we may choose as a measure of goodness of fit the ratio

$$\frac{\lambda_{q+1} + \cdots + \lambda_p}{\lambda_1 + \cdots + \lambda_p} . \tag{3.10}$$

The choice of q, then depends on the smallness of (3.10) or the largeness of the ratio

$$\frac{\lambda_1 + \cdots + \lambda_q}{\lambda_1 + \cdots + \lambda_p} . \tag{3.11}$$

It may happen that the overall measure (3.11) has a small value but the configuration of the points as a whole is distorted in the projected space due to some isolated points being far away from the best fitting space. The examination of such isolated points is of interest in practical work and may provide an interpretation of the nature of heterogeneity in the data. For this purpose it is necessary to compute the length of the perpendicular of each point on the best fitting space. The square of the perpendicular from $\underline{X}_i$, the i^{th} point is

$$d_i^2 = (\underline{X}_i - \underline{\overline{X}})'(\underline{X}_i - \underline{\overline{X}}) - [\underline{P}_1'(\underline{X}_i - \underline{\overline{X}})]^2 - \cdots - [\underline{P}_q'(\underline{X}_i - \underline{\overline{X}})]^2$$

$$i = 1, \ldots, n .$$

A comparison of the values $d_1^2, \ldots, d_n^2$ will enable us to find out the outliers if any.

4. Other interpretations of the transformation (3.8).

The problem of representing the points (3.1) in a lower dimensional space may be posed as the determination of a transformation matrix $\underline{T}$

of order $p \times q$, and rank q transforming a p-vector $\underline{X}$ into a q-vector $\underline{Y}$, $\underline{Y} = \underline{T}'\underline{X}$, and satisfying some optimum properties. The transformation from $\underline{Y}$ to $\underline{X}$ is not one to one when $q < p$ and consequently certain properties of the configuration of the points in the original space (which is invariant when $q = p$) are altered. Our aim is to choose $\underline{T}$ which preserves the configuration of the points to the maximum possible extent. We shall show that some intuitive measures of closeness between the configurations in the original space and the transformed subspace, when minimized with respect to the transformation matrix $\underline{T}$ lead to the transformation (3.8) as the optimum.

Let us observe that the columns of $\underline{T}$ can be chosen to be orthonormal, without loss of generality.

Maximizing the sum of squares of the distances. The sum of squares of the distances between all possible pairs of points in the p-space is

$$(4.1) \qquad \sum\sum_{i,j=1}^{n} (\underline{X}_i-\underline{X}_j)'(\underline{X}_i-\underline{X}_j) = n \text{ trace } \underline{\Sigma},$$

while the corresponding expression in the q-space is

$$(4.2) \qquad \sum\sum_{i,j=1}^{n} (\underline{Y}_i-\underline{Y}_j)'(\underline{Y}_i-\underline{Y}_j) = n \text{ trace } T' \, \Sigma \, T$$

$$= n(\underline{T}_1' \, \underline{\Sigma} \, \underline{T}_1 + \cdots + \underline{T}_q' \, \underline{\Sigma} \, \underline{T}_q)$$

where $T_1,\ldots,T_q$ are the columns of $\underline{T}$. It is seen that (4.2) < (4.1) and (4.2) = (4.1) when (not necessarily only when) $p = q$. Let us choose $\underline{T}$ such that (4.2) is a maximum. Applying the result (2.4), we find that the maximum is attained when $\underline{T}_i^* = \underline{P}_i$, $i=1,\ldots,q$, which is the solution obtained in section 3.

Closest fit to the distances and the angles of the lines joining the points to the center of gravity. In the p-space the vectors joining the points to their center of gravity are represented by the matrix $\underset{\sim}{X}_d$ as defined in (3.3). It may be seen that in the matrix

$$\underline{\Lambda} = \underset{\sim}{X}'_d \underset{\sim}{X}_d \tag{4.3}$$

the i^{th} diagonal entry is the square of the distance of the i^{th} point from the center of gravity and the (i,j) entry divided by the square root of the i^{th} and j^{th} entries is the cosine of the angle between the lines joining the i^{th} and j^{th} points to the center of gravity. Thus the matrix $\underline{\Lambda}$ represents the distances and the angles between the lines joining the points to their center of gravity. In the q-space, under the transformation $\underline{Y} = \underline{T}'\underline{X}$ the matrix corresponding to (4.3) is

$$\underline{B} = \underset{\sim}{Y}'_d \underset{\sim}{Y}_d = \underset{\sim}{X}'_d \underline{T}\,\underline{T}' \underset{\sim}{X}_d \ . \tag{4.4}$$

We wish to determine $\underline{T}$ such that the matrix $\underline{B}$ is as close as possible to $\underline{\Lambda}$. We may measure the closeness of $\underline{B}$ and $\underline{\Lambda}$ by the Euclidean norm $\|\underline{\Lambda}-\underline{B}\|$, which is the square root of the sum of squares of the elements in $(\underline{\Lambda}-\underline{B})$. Then the problem is one choosing $\underline{B}$ such that $\|\underline{\Lambda}-\underline{B}\|$ is a minimum. An application of the result (2.6) gives the optimum choice of $\underline{B}$ as

$$\underset{\sim}{X}'_c \underline{T}\,\underline{T}' \underset{\sim}{X}_d = \lambda_1 \underline{Q}_1 \underline{Q}'_1 + \dots + \lambda_q \underline{Q}_q \underline{Q}'_q \tag{4.5}$$

where $\underline{Q}_1, \dots, \underline{Q}_q$ are the eigen vectors of $\underline{\Lambda} = \underset{\sim}{X}'_d \underset{\sim}{X}_d$. The optimum choice of T is obtained from (4.5) as

$$(4.6) \qquad \begin{aligned} \underset{\sim}{X}_d \, \underline{T}^* &= (\sqrt{\lambda_1}\underline{Q}_1 \; \cdots \; \sqrt{\lambda_q}\underline{Q}_q) \\ &= (\underset{\sim}{X}'_d \, \underline{P}_1 \; \cdots \; \underset{\sim}{X}'_d \, \underline{P}_q) \end{aligned}$$

using the result of (iv), section 2. The choice

$$(4.7) \qquad \underline{T}^* = (\underline{P}_1 \cdots \underline{P}_q)$$

satisfies the equation (4.6). Thus $\underline{T}^*_i = \underline{P}_i$, where $\underline{T}^*_i$ is the i^{th} column of $\underline{T}^*$, which is the solution obtained in section 3. This means that the representation of the points in a q dimensional space as in (3.8) preserves the distances and angles of the configuration to a large extent.

5. Reduction in the number of points.

In sections 3 and 4, a reduction in the number of dimensions of the space is secured through a transformation of the type $\underline{Y} = \underline{T}'\underline{X}$ by choosing an optimum $\underline{T}$. We now consider the dual problem of reducing the number of points to a few typical points keeping the number of dimensions the same. For this purpose we consider the matrix $\underset{\sim}{X}_d$ which represents the points as measured from the center of gravity. The reduction in the number of points is secured by a transformation of the type $\underset{\sim}{Y}_d = \underset{\sim}{X}_d\underline{M}$ where $\underline{M}$ is n × q matrix.

Reduction providing the closest fit to the total dispersion matrix. The total dispersion matrix based on the reduced points is

$$(5.1) \qquad \underset{\sim}{Y}_d\underset{\sim}{Y}'_d = \underset{\sim}{X}_d \, \underline{M} \, \underline{M}' \, \underset{\sim}{X}'_d$$

while that based on all points is $\underline{\Sigma}$. We wish to choose $\underset{\sim}{Y}_d$ in such a way that

$$\|\underset{\sim}{\Sigma}-\underset{\sim}{Y}_d\,\underset{\sim}{Y}_d'\| \tag{5.2}$$

is a minimum. Applying the result (2.6), the minimum of (5.2) is attained when

$$\underset{\sim}{Y}_d\,\underset{\sim}{Y}_d' = \lambda_1\underline{P}_1\underline{P}_1' + \cdots + \lambda_q\underline{P}_q\underline{P}_q'$$

or when

$$\underset{\sim}{Y}_d = (\sqrt{\lambda_1}\underline{P}_1 \cdots \sqrt{\lambda_q}\underline{P}_q)\ . \tag{5.3}$$

Thus the typical points summarizing the deviations from the center of gravity of the original points are the eigen vectors scaled by the square roots of the corresponding eigen values.

To determine the transformation matrix (if necessary) let us observe that, using result (iv) of section 2,

$$\begin{aligned} \underset{\sim}{Y}_d &= (\sqrt{\lambda_1}\underline{P}_1 \cdots \sqrt{\lambda_q}\underline{P}_q) \\ &= (\underset{\sim}{X}_d\underline{Q}_1 \cdots \underset{\sim}{X}_d\underline{Q}_q)\ , \end{aligned} \tag{5.4}$$

where $\underline{Q}_1,\ldots,\underline{Q}_q$ are the first q eigen vectors of the matrix $\underset{\sim}{X}_d'\,\underset{\sim}{X}_d$. But (5.4) is the same as $\underset{\sim}{X}_d\,\underline{M}^*$ where $\underline{M}^*$ is the matrix with $\underline{Q}_1,\ldots,\underline{Q}_q$ as its columns.

An interpretation of the typical points. In section 3, it is shown that the best fitting q-space is specified by $\underline{\bar{X}}$, the center of gravity and the eigen vectors $\underline{P}_1,\ldots,\underline{P}_q$. Any point in this space referred to the p original axes can be represented by a vector of the form

$$\underline{\bar{X}} + a_1\underline{P}_1 + \cdots + a_q\underline{P}_q \tag{5.5}$$

where $a_1,\ldots,a_q$ are arbitrary. By a change in scale, (5.5) can be written as

$$\underline{\bar{X}} + b_1\sqrt{\lambda_1}\,\underline{P}_1 + \cdots + b_q\sqrt{\lambda_q}\,\underline{P}_q \tag{5.6}$$

involving the typical points $\sqrt{\lambda_1}\,\underline{P}_1, \ldots, \sqrt{\lambda_q}\,\underline{P}_q$. Let $\underline{X}_i^{(g)}$ be the projection of $\underline{X}_i$ on this subspace. Then there exist constants $b_1, \ldots, b_q$ such that

$$\underline{X}_i^{(g)} = \underline{\bar{X}} + b_{1i}\sqrt{\lambda_1}\,\underline{P}_1 + \cdots + b_{qi}\sqrt{\lambda_q}\,\underline{P}_q \; . \tag{5.7}$$

Multiplying both sides of (5.7) by $\underline{P}_j'$ and observing that $\underline{P}_j'\underline{X}_i^{(g)} = \underline{P}_j'\underline{X}_i$, we obtain the value of b_{ji} as

$$b_{ji} = \lambda^{-1/2}\underline{P}_j'(\underline{X}_i - \underline{\bar{X}}) \; . \tag{5.8}$$

With such a choice of b_{ji}

$$\underline{X}_i^{(g)} = \underline{\bar{X}} + b_{1i}\sqrt{\lambda_1}\,\underline{P}_1 + \cdots + b_{qi}\sqrt{\lambda_q}\,\underline{P}_q \, , \qquad i=1,\ldots,n \; . \tag{5.9}$$

The point $\underline{X}_i^{(g)}$ so determined may be called the graduated or the fitted point of $\underline{X}_i$. If $\underline{X}_i - \underline{X}_i^{(g)}$ is small for each i, the result (5.9) shows that the average point and the typical points

$$\underline{\bar{X}} \, , \sqrt{\lambda_1}\,\underline{P}_1, \ldots, \sqrt{\lambda_q}\,\underline{P}_q$$

form an approximate basis of the n observed points

$$\underline{X}_i \sim \underline{\bar{X}} + b_{1i}\sqrt{\lambda_1}\,\underline{P}_1 + \cdots + b_{qi}\sqrt{\lambda_q}\,\underline{P}_q \qquad i=1,\ldots,n \tag{5.10}$$

The representation (5.10) admits a statistical interpretation under the following model for the observed points X_i:

$$(5.11) \qquad \underline{X}_i = \underline{\mu} + \beta_{1i}\underline{\pi}_1 + \cdots + \beta_{qi}\underline{\pi}_q + \underline{\epsilon}_i$$

where $\underline{\mu}$, $\underline{\pi}_1, \ldots, \underline{\pi}_q$ are fixed vectors, $\beta_{1i}, \ldots, \beta_{qi}$ are constants specific to the i^{th} point and $\underline{\epsilon}_i$ is a vector of random errors. The model (5.11) implies that any observed point when corrected for errors of measurement can be expressed as linear combination of $(q+1)$ basic vectors only. It can be shows that, when the dispersion matrix of $\underline{\epsilon}_i$ is of the form $\sigma^2 I$, consistent estimators of $\underline{\mu}$, $\underline{\pi}_i$ and β_{ji} are provided by $\underline{\bar{X}}$, $\sqrt{\lambda_i}\,\underline{P}_i$ and b_{ji} as defined in (5.10). Under the additional condition that the p components of $\underline{\epsilon}_i$ have independent normal distributions, $\underline{\bar{X}}$, $\sqrt{\lambda_i}\,\underline{P}_i$ and b_{ji} can be shown to be maximum likelihood estimators. Such a statistical interpretation is not true when the dispersion matrix of $\underline{\epsilon}_i$ is not of the form $\sigma^2 \underline{I}$. This is true only when the errors of measurements of the components of each point $\underline{X}_i$ are uncorrelated and have the same variance.

The typical points $\sqrt{\lambda_1}\,\underline{P}_1, \ldots, \sqrt{\lambda_q}\,\underline{P}_q$ used in the representation (5.10) may not admit any physical interpretation. But in special situations it may be possible to characterize each typical point as indicating some aspect of the measurements as a whole (see Simonds, 1964).

6. Fitting a subspace when the points are in a oblique space.

In the analysis of section 3, it is implicitly assumed that the original points are represented in a p-dimensional Euclidean space with orthogonal axes. There are, however, situations where oblique axes are chosen to represent the coordinates. In such a case, the distance between any two points $\underline{X}_i$, $\underline{X}_j$ is a quadratic form

$$(\underline{X}_i-\underline{X}_j)'\underline{\Gamma}^{-1}(\underline{X}_i-\underline{X}_j) \tag{6.1}$$

using a positive definite matrix $\underline{\Gamma}$. The scatter of points analogous to (3.10) is

$$\sum_{i,j=1}^{n}\sum (\underline{X}_i-\underline{X}_j)'\underline{\Gamma}^{-1}(\underline{X}_i-\underline{X}_j) \ . \tag{6.2}$$

Let us transform $\underline{X}$ to $\underline{Y}$ with q coordinates

$$\underline{Y}' = (\underline{L}_1'\underline{X},\ldots,\underline{L}_q'\underline{X})$$

where $\underline{L}_i' \ \underline{\Gamma} \ \underline{L}_i = 1$ and $\underline{L}_i' \ \underline{\Gamma} \ \underline{L}_j = 0$, $i \neq j$. The scatter of points in the q-space with $\underline{L}_1,\ldots,\underline{L}_q$ as orthogonal axes is

$$\sum\sum (\underline{Y}_i-\underline{Y}_j)'(\underline{Y}_i-\underline{Y}_j)=n(\underline{L}_1' \ \underline{\Sigma} \ \underline{L}_1 + \cdots + \underline{L}_q' \ \underline{\Sigma} \ \underline{L}_q). \tag{6.3}$$

An application of the result (2.10) shows that (6.3) is a maximum when $\underline{L}_i = \underline{P}_i$, $i=1,\ldots,q$, where $\underline{P}_1,\ldots,\underline{P}_q$ are the first q eigen vectors associated with the determinental equation $|\underline{\Sigma}-\underline{\Gamma}| = 0$. The adequacy of a q dimensional fit is judged by the largeness of the ratio

$$\frac{\lambda_1 + \cdots + \lambda_q}{\lambda_1 + \cdots + \lambda_p} \tag{6.4}$$

where $\lambda_1,\ldots,\lambda_p$ are the roots of $|\underline{\Sigma}-\lambda\underline{\Gamma}| = 0$.

7. Principal components of a vector random variable.

The concept of principal component analysis as applied to a random variable is due to Hotelling (1933, 1935, 1936a).

Let us consider a p dimensional random variable $\underline{X}$ with mean $\underline{\mu}$ and dispersion matrix $\underline{\Sigma}$. Consider a transformation, $\underline{Y} = \underline{T}'\underline{X}$, where $\underline{T}$ is a $p \times q$ matrix, so that $\underline{Y}$ is q dimensional. When $q < p$, there is loss of information in replacing $\underline{X}$ by $\underline{Y}$. For a given q, we wish to determine $\underline{T}$ such that there is minimum loss.

In defining a loss function we may be guided by the question as to what extent we can predict $\underline{X}$ knowing $\underline{Y}$. The predictive efficiency of $\underline{Y}$ for $\underline{X}$ depends on the residual dispersion matrix of $\underline{X}$ after subtracting its best linear predictor in terms of $\underline{Y}$. Now the joint dispersion matrix of $\underline{X}$ and $\underline{Y}$ is

$$(7.1) \qquad \begin{pmatrix} \underline{\Sigma} & \underline{\Sigma}\,\underline{T} \\ \underline{T}'\underline{\Sigma} & \underline{T}'\,\underline{\Sigma}\,\underline{T} \end{pmatrix}$$

and the residual dispersion matrix is

$$(7.2) \qquad \underline{\Sigma}-\underline{\Sigma}\,\underline{T}(\underline{T}'\,\underline{\Sigma}\,\underline{T})^{-1}\underline{T}'\,\underline{\Sigma}\ .$$

The smaller the values of the elements in (7.2), the greater is the predictive efficiency. We consider two overall measures, the trace and the Euclidean norms of (7.2),

$$(7.3) \qquad \text{(a)}\quad \text{Trace}(\underline{\Sigma}-\underline{\Sigma}\,\underline{T}(\underline{T}'\,\underline{\Sigma}\,\underline{T})^{-1}\underline{T}'\underline{\Sigma})$$

$$(7.4) \qquad \text{(b)}\quad \|\underline{\Sigma}-\underline{\Sigma}\,\underline{T}(\underline{T}'\,\underline{\Sigma}\,\underline{T})^{-1}\underline{T}'\underline{\Sigma}\|\ .$$

We shall show that any one of these measures when minimized leads to the same choice of $\underline{T}$ as the matrix of the first q eigen vectors of $\underline{\Sigma}$.

Let us observe that the column vectors $\underline{T}_1, \ldots, \underline{T}_q$ of $\underline{T}$ can be chosen, without loss of generality, to satisfy the conditions

$$\underline{T}_i' \, \underline{\Sigma} \, \underline{T}_j = 0 \, , \quad i \neq j \tag{7.5}$$

which imply that the components of the transformed variable are uncorrelated.

Consider the measure (7.3)

$$\begin{aligned} & \mathrm{Trace}(\underline{\Sigma} - \underline{\Sigma} \, \underline{T}(\underline{T}' \, \underline{\Sigma} \, \underline{T})^{-1} \underline{T}' \underline{\Sigma}) \\ & = \mathrm{Trace} \, \underline{\Sigma} - \mathrm{Trace}(\underline{T}' \, \underline{\Sigma} \, \underline{T})^{-1} \underline{T}' \underline{\Sigma} \, \underline{\Sigma} \, \underline{T}) \\ & = \mathrm{trace} \, \underline{\Sigma} - \left(\frac{\underline{T}_1' \underline{\Sigma} \, \underline{\Sigma} \, \underline{T}_1}{\underline{T}_1' \, \underline{\Sigma} \, \underline{T}_1} + \cdots + \frac{\underline{T}_q' \underline{\Sigma} \, \underline{\Sigma} \, \underline{T}_q}{\underline{T}_q' \, \underline{\Sigma} \, \underline{T}_q} \right) \end{aligned} \tag{7.6}$$

using the conditions (7.5). Minimizing (7.6) is the same as maximizing

$$\frac{\underline{T}_1' \underline{\Sigma} \, \underline{\Sigma} \, \underline{T}_1}{\underline{T}_1' \, \underline{\Sigma} \, \underline{T}_1} + \cdots + \frac{\underline{T}_q' \underline{\Sigma} \, \underline{\Sigma} \, \underline{T}_q}{\underline{T}_q' \, \underline{\Sigma} \, \underline{T}_q} \, .$$

Applying the result (2.10) the optimum choice of $\underline{T}_i$ is the i^{th} eiger vector $\underline{\Sigma} \, \underline{\Sigma}$ with respect to $\underline{\Sigma}$, which is the same as $\underline{P}_i$ the i^{th} eigen vector of $\underline{\Sigma}$. The minimum value of (7.6) is then

$$\lambda_{q+1} + \cdots + \lambda_p \, . \tag{7.7}$$

The sum of the smallest $p-q$ eigen values of $\underline{\Sigma}$.

Consider the problem of minimizing (7.6),

$$\| \underline{\Sigma} - \underline{\Sigma} \, \underline{T}(\underline{T}' \, \underline{\Sigma} \, \underline{T})^{-1} \underline{T}' \, \underline{\Sigma} \| \, . \tag{7.8}$$

The result (2.6) shows that the minimum of (7.7) is attained when

$$\underline{\Sigma} \, \underline{T}(\underline{T}' \, \underline{\Sigma} \, \underline{T})^{-1} \underline{T} \, \underline{\Sigma} = \lambda_1 \, \underline{P}_1 \, \underline{P}_1' + \cdots + \lambda_q \, \underline{P}_q \, \underline{P}_q' \tag{7.9}$$

and it is easy to verify that (7.9) holds when $T_i^* = P_i$. The minimum value of (7.8) is

$$\lambda_{q+1}^{2} + \cdots + \lambda_{p}^{2} \tag{7.10}$$

where $\lambda_{q+1}, \ldots, \lambda_p$ are as defined in (7.8).

The transformed variables $\underline{P}_1'\underline{X}, \ldots, \underline{P}_q'\underline{X}$ are called the first q principal components of the random variable $\underline{X}$ and are interpreted variously. The usual interpretation is as follows. Under a complete orthogonal transformation $\underline{Y} = \underline{O}'\underline{X}$, i.e., from a p dimensional variable to another p dimensional variable, the trace of the dispersion matrix, which is the sum of the variances of the variables, remains invariant. Thus

$$\text{Trace } \underline{\Sigma} = \text{Trace } \underline{O}'\underline{\Sigma}\,\underline{O} = \underline{O}_1'\,\underline{\Sigma}\,\underline{O}_1 + \cdots + \underline{O}_p'\,\underline{\Sigma}\,\underline{O}_p \tag{7.11}$$

where $\underline{O}'\,\underline{\Sigma}\,\underline{O}$ is the dispersion matrix of $\underline{Y}$ and $\underline{O}_i$ is the i^{th} column vector of $\underline{O}$. The total of the variances of the first q variables in $\underline{Y}$ is

$$\underline{O}_1'\underline{\Sigma}\,\underline{O}_1 + \cdots + \underline{O}_q'\underline{\Sigma}\,\underline{O}_q \leq \text{trace } \underline{\Sigma} \,. \tag{7.12}$$

It is seen from (2.6) that the optimum choice of $\underline{O}_i$ is $\underline{P}_i$ for (7.12) to be a maximum, i.e., the first q transformed variables are $\underline{P}_1'\underline{X}, \ldots, \underline{P}_q'\underline{X}$, are the q principal components of $\underline{X}$.

The maximum value of (7.12) is $\lambda_1 + \cdots + \lambda_q$ while the total of the variances of the variables is $\lambda_1 + \cdots + \lambda_p$, in which case the first q principal components are said to explain $100(\lambda_1 + \cdots + \lambda_q)/(\lambda_1 + \cdots + \lambda_p)$ percent of the total variance.

The interpretation of the principal components as the best predictors of $\underline{X}$ suggests extensions of the principal component analysis in many directions. which are considered in the later sections.

8. Principal components of instrumental variables.

Let $\underline{X}$ be the vector of p main variables, and $\underline{Z}$ the vector of m instrumental variables. In theory $\underline{Z}$ may include some or all the elements of $\underline{X}$. Denote the joint dispersion matrix of $(\underline{X},\underline{Z})$ by

$$\begin{pmatrix} \underline{\Sigma} & \underline{\Theta} \\ \underline{\Theta}' & \underline{\Gamma} \end{pmatrix} . \tag{8.1}$$

We wish to replace $\underline{Z}$ by a q dimensional random variable $\underline{Y} = \underline{M}'\underline{Z}$ in such a way that the predictive efficiency of $\underline{Y}$ for $\underline{X}$ is a maximum, (Rao, 1962a). The dispersion matrix of $(\underline{X},\underline{Y})$ is

$$\begin{pmatrix} \underline{\Sigma} & \underline{\Theta}\,\underline{M} \\ \underline{M}'\underline{\Theta}' & \underline{M}'\underline{\Gamma}\,\underline{M} \end{pmatrix} \tag{8.2}$$

and the residual dispersion matrix of $\underline{X}$ subtracting its best linear predictor in terms of $\underline{Y}$ is

$$\underline{\Sigma} - \underline{\Theta}\,\underline{M}(\underline{M}'\underline{\Gamma}\,\underline{M})^{-1}\underline{M}'\underline{\Theta} \quad . \tag{8.3}$$

As in section 7, we may consider the two measures of predictive efficiency of $\underline{Y}$

$$\text{(a)}\quad \text{Trace}(\underline{\Sigma}\text{-}\underline{\Theta}\,\underline{M}(\underline{M}'\,\underline{\Gamma}\,\underline{M})^{-1}\underline{M}'\,\underline{\Theta}') \tag{8.4}$$

$$\text{(b)}\quad \|\underline{\Sigma}\text{-}\underline{\Theta}\,\underline{M}(\underline{M}'\,\underline{\Gamma}\,\underline{M})^{-1}\underline{M}'\,\underline{\Theta}'\|. \tag{8.5}$$

Unfortunately, the solution seems to depend on which measure is chosen for minimization.

Minimizing (8.4) is the same as maximizing

$$
\begin{aligned}
&\text{Trace}\ \underline{\Theta}\ \underline{M}(M'\ \Gamma\ M)^{-1}M'\ \Theta' \\
(8.6)\qquad &= \text{trace}\ (M'\ \Gamma\ M)^{-1}M'\ \Theta'\ \Theta\ M \\
&= \frac{\underline{M}_1'\ \underline{\Theta}'\underline{\Theta}\ \underline{M}_1}{\underline{M}_1'\ \underline{\Gamma}\ \underline{M}_1} + \dots + \frac{\underline{M}_q'\ \underline{\Theta}'\underline{\Theta}\ \underline{M}_q}{\underline{M}_q'\ \underline{\Gamma}\ \underline{M}_q}
\end{aligned}
$$

assuming that $\underline{M}_i\ \underline{\Gamma}\ \underline{M}_j = 0,\ i \neq j$, without loss of generality. Applying the result (2.10), the best choice of $\underline{M}_1, \dots, \underline{M}_q$ is the set of the first q eigen vectors of the matrix $\underline{\Theta}'\underline{\Theta}$ with respect to Γ, i.e., associated with the determinental equation

$$
(8.7)\qquad |\underline{\Theta}'\underline{\Theta} - \lambda\ \underline{\Gamma}| = 0\ .
$$

It may be observed that when $\underline{X}$ is a proper subset of $\underline{Z}$, the optimum choice of $\underline{Y}$ consists of the first q principal components of $\underline{X}$ alone.

The present analysis is different from that of canonical correlations and canonical variates as developed by Hotelling (1936) for studying the association between two vector random variables.

There seems to be no simple method of minimizing the measure (8.5). The solution would be, however, different from that obtained by minimizing (8.4). The relative merits of the two solutions are worth exploring.

9. Weighted principal components.

The measures of predictive efficiency (7.3), (7.4), (8.4) and (8.5) are not invariant under change of scales of the elements of $\underline{X}$. This is an undesirable feature as different solutions to principal components can be found by choosing different scales. The difficulty can be overcome by defining a loss function associated with the prediction of each element of $\underline{X}$ by any chosen set of q linear functions $\underline{Y}$ of the instrumental variable $\underline{Z}$. A simple loss function for the prediction of $\underline{X}_i$, the i^{th} element of $\underline{X}$, may be chosen as w_i^2 times the residual variance, where w_i^2, $i=1,\ldots,p$, are assigned quantities. Denoting by W, the diagonal matrix of the elements $w_1,\ldots,w_p$ the total loss in the prediction of $\underline{X}$ is the weighted trace

$$(9.1) \qquad \text{trace } \underline{W}\,\underline{\Sigma}\,\underline{W} - \text{trace } \underline{W}\,\underline{\Theta}\,\underline{M}(\underline{M}'\,\underline{\Gamma}\,\underline{M})^{-1}\underline{M}'\,\underline{\Theta}'\,\underline{W} \ .$$

The problem is one of maximizing

$$(9.2) \qquad \text{trace } \underline{W}\,\underline{\Theta}\,\underline{M}\,\underline{M}'\,\underline{\Theta}'\,\underline{W} = \text{trace } \underline{M}'\,\underline{\Theta}'\,\underline{W}^2\,\underline{\Theta}\,\underline{M} \ ,$$

choosing $\underline{M}'\,\underline{\Gamma}\,\underline{M} = I$, without loss of generality. The best choice of $\underline{M}$ is the matrix of the first q eigen vectors associated with the determinental equation

$$(9.3) \qquad |\underline{\Theta}'\,\underline{W}^2\,\underline{\Theta} - \lambda\underline{\Gamma}| = 0 \ .$$

In the special case when the instrumental variable is the same as (or contains) the main variable, the equation (9.3) reduces to

$$(9.4) \qquad |\,\underline{\Sigma}\,\underline{W}^2\underline{\Sigma} - \lambda\underline{\Sigma}| = 0$$

which is equivalent to

$$(9.5) \qquad |\underline{\Sigma} - \lambda\underline{W}^{-2}| = 0 \ .$$

10. Principal components with restrictions on individual variances.

In the previous sections the predictive efficiency of a random variable $\underline{Y}$ for predicting $\underline{X}$ is judged by the sum (or weighted sum) of residual variances for the different components of $\underline{X}$. An optimum choice of $\underline{Y}$ (in a given class) was made by maximizing the predictive efficiency, i.e., by minimizing the sum of residual variances. But it may so happen that for such an optimum choice of $\underline{Y}$, the residual variances for some elements of $\underline{X}$ are above certain desired levels. In such a case it may be necessary to impose some restrictions on the individual variances and maximize the overall predictive efficiency.

Let $\underline{Y}$ be the vector of q linear functions, $\underline{L}_1'\underline{X},\ldots,\underline{L}_q'\underline{X}$, of $\underline{X}$, where $\underline{L}_i'\underline{\Sigma}\,\underline{L}_i = 1$, $\underline{L}_i'\underline{\Sigma}\,\underline{L}_j = 0$, $i \neq j$. The residual variance in predicting X_i, the i^{th} element of $\underline{X}$, by $\underline{Y}$ is

$$(10.1) \qquad \sigma_{ii}-(\underline{L}_1'\underline{\Sigma}_i)^2 - \cdots - (L_q'\Sigma_i)^2$$

where $\underline{\Sigma}_i$ is the i^{th} column vector of $\underline{\Sigma}$. We impose the conditions

$$(10.2) \qquad \begin{aligned} &\sigma_{ii}-(\underline{L}_1'\underline{\Sigma}_i)^2 - \cdots - (\underline{L}_q'\underline{\Sigma}_i)^2 \leq e_i^2 \text{ (given)} \\ \text{or} \quad &(\underline{L}_1'\underline{\Sigma}_i)^2+\cdots+(\underline{L}_q'\underline{\Sigma}_i)^2 \geq \sigma_{ii}-e_i^2 = (\delta_i^2 \text{ (say)}) \\ &i=1,\ldots,p\ . \end{aligned}$$

The sum of squares of residual variances is

$$(10.3) \qquad \Sigma\sigma_{ii} - (\underline{L}_1'\underline{\Sigma}\,\underline{\Sigma}\,\underline{L}_1+\cdots+\underline{L}_q'\underline{\Sigma}\,\underline{\Sigma}\,\underline{L}_q)\ .$$

The problem is one of maximizing

$$(10.4) \qquad \underline{L}_1'\underline{\Sigma}\,\underline{\Sigma}\,\underline{L}_1+\cdots+\underline{L}_q'\underline{\Sigma}\,\underline{\Sigma}\,\underline{L}_q$$

subject to the conditions

$$(10.5) \qquad \begin{cases} \underline{L}_i'\underline{\Sigma}\,\underline{L}_i = 1\ , \quad i=1,\ldots,q \\ \underline{L}_i'\underline{\Sigma}\,\underline{L}_j = 0\ , \quad i\neq j\ , \\ (\underline{L}_1'\Sigma_i)^2+\cdots+(\underline{L}_q'\Sigma_i)^2 \geq \delta_i^2\ , \ i=1,\ldots,p\ . \end{cases}$$

We are thus led to a very complicated problem in non-linear programming.

When $q = 1$, there is only a single linear function $\underline{L}_1'\underline{X}$ to be determined. The expression to be maximized is

$$(10.6) \qquad \underline{L}_1' \, \underline{\Sigma} \, \underline{\Sigma} \, \underline{L}_1$$

subject to the conditions

$$(10.7) \qquad \underline{L}_1' \, \underline{\Sigma} \, \underline{L}_1 = 1 \quad \text{and} \quad (\underline{L}_1' \, \Sigma_i)^2 \geq \delta_i^2 \, , \; i=1,\ldots,p \, .$$

Making the substitution $\underline{\Sigma}\,\underline{L}_1 = \underline{U}$ or $\underline{L}_1 = \underline{\Sigma}^{-1}\underline{U}$, we need only maximize

$$(10.8) \qquad \underline{U}'\underline{U}$$

subject to the conditions

$$(10.9) \qquad \underline{U}' \, \underline{\Sigma}^{-1} \, \underline{U} = 1 \quad \text{and} \quad U_i^2 \geq \delta_i^2 \, , \; i=1,\ldots,p \, .$$

The solution is complicated even in this simple case. It has been shown elsewhere (Rao, 1962b, 1964) how to obtain a solution for small values of p. The general problem awaits solution.

11. Principal components of X uncorrelated with the instrumental variable Z.

In section 8, the instrumental variable $\underline{Z}$ is analyzed into principal components on the basis of their predictive efficiency for elements of a variable $\underline{X}$. In some problems, it is of interest to determine the principal components of $\underline{X}$ in the class of linear functions of $\underline{X}$ uncorrelated with instrumental variables. The problem may be stated as one of determining q linear functions $\underline{L}_1'\underline{X}, \ldots, \underline{L}_q'\underline{X}$ such that

$$(11.1) \qquad V(\underline{L}_1'\underline{X}) + \cdots + V(\underline{L}_q'\underline{X}) = \underline{L}_1' \, \underline{\Sigma} \, \underline{L}_1 + \cdots + \underline{L}_q' \, \underline{\Sigma} \, \underline{L}_q$$

is a maximum subject to the conditions

$$(11.2)\qquad \begin{cases} \underline{L}_i'\underline{L}_i = 1 \ , \ \underline{L}_i'\underline{L}_j = 0 \\ \\ \mathrm{Cov}(\underline{L}_i'\underline{X},\underline{Z}) = 0 \ , \ i=1,\ldots,q \ . \end{cases}$$

As before, let the dispersion matrix of (X,Z) be

$$(11.3)\qquad \begin{pmatrix} \underline{\Sigma} & \underline{\Theta} \\ \underline{\Theta}' & \underline{\Gamma} \end{pmatrix} .$$

In terms of the elements of the matrix (11.3), the problem is one of maximizing

$$(11.4)\qquad \underline{L}_1' \ \underline{\Sigma} \ \underline{L}_1 + \cdots + \underline{L}_q' \ \underline{\Sigma} \ \underline{L}_q$$

subject to the conditions

$$(11.5)\qquad \begin{aligned} &\underline{L}_i'\underline{L}_i = 1 \ , \ \underline{L}_i'\underline{L}_j = 0 \ , \ i \neq j \\ &\underline{L}_i' \ \underline{\Theta} = 0 \ , \\ &i,j=1,\ldots,q \ . \end{aligned}$$

An application of the result (2.12) shows that the maximum of (11.4) is attained when $\underline{L}_1,\ldots,\underline{L}_q$ are the first q right eigen vectors of the matrix

$$(11.6)\qquad (\underline{I}-\underline{\Theta}(\underline{\Theta}'\underline{\Theta})^{-1}\underline{\Theta}')\underline{\Sigma} \ .$$

The principal components so determined may be useful in the following situation.

Suppose we have a vector variable $\underline{X}$ representing p 'economic transactions' measured in terms of a common unit (like the cash value). The observed values of $\underline{X}$ over time constitute a multiple time series. Some years ago, Stone (1947) considered the problem of isolating linear functions of $\underline{X}$ which have an intrinsic economic significance from those which represent trend with time and those which measure random errors. For this purpose, he computed the dispersion matrix of the p-variables considering the observations at different points of time as repeated values of $\underline{X}$ and extracted the principal components of $\underline{X}$ from the estimated dispersion matrix, without any reference to the time factor. The problem was then posed as that of interpreting the dominant principal components which accounted for a high percentage of the total variance. The first principal component with the largest variance was interpreted as representing linear trend and the rest were interpreted in economic terms.

It is not clear why the trend, even if it is linear, should be reflected in the first principal component only. Then there is the difficulty of choosing more than one principal component to explain the trend, if it is non-linear in time.

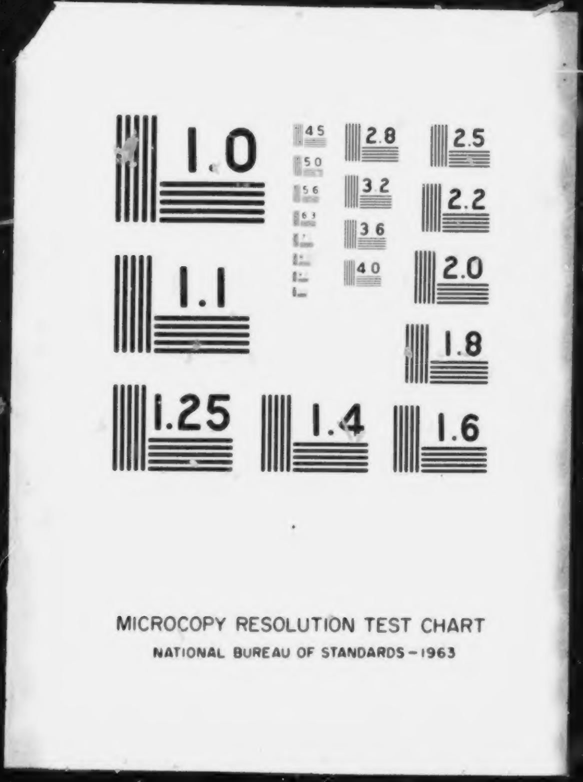
To apply the present analysis, we consider an instrumental variable $\underline{Z}$ consisting of orthogonal functions of time, $\phi_1(t), \phi_2(t), \ldots,$ representing the first, second,... degree polynomial trend with respect to time. Then we obtain the principal components of $\underline{X}$ in the class of linear functions not showing any trend, which may then be interpreted in economic terms.

12. Size and shape factors.

Biologists use functions of measurements representing size and shape of organisms in studying differences within and between groups. Size and shape are not, however, well defined concepts and their measurement will be arbitrary to some extent. Let us examine some functions proposed for this purpose.

4 OF 6

ED
003255


1.0
4.5
5.0
5.6
6.3
2.8
3.2
3.6
4.0
2.5
2.2
2.0
1.1
1.8
1.25
1.4
1.6

MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963

Penrose (1947) defines size as the linear function

$$(12.1)\qquad \frac{X_1}{\sigma_1} + \dots + \frac{X_p}{\sigma_p}$$

where σ_i is the standard deviation of X_i and shape as any linear function

$$(12.2)\qquad c_1 \frac{X_1}{\sigma_1} + \dots + c_p \frac{X_p}{\sigma_p}$$

with $c_1 + \dots + c_p = 0$. In situations where X_1, X_2, X_3 correspond, in some sense, to length, width and height of an organism, Mosimann (1950) proposed the product

$$(12.3)\qquad X_1 X_2 X_3$$

as a volumetric or a ponderal definition of size. It may be noted that a function of the type

$$(12.4)\qquad X_1^{\beta_1} X_2^{\beta_2} X_3^{\beta_3}$$

studied by Rao and Shaw (1948) provides a better index of volume for a suitable choice of the exponents $\beta_1, \beta_2, \beta_3$. In terms of the logarithms of the measurements the size function (12.4) has the linear form

$$(12.5)\qquad \beta_1 \log X_1 + \beta_2 \log X_2 + \beta_3 \log X_3 .$$

It has been suggested by Jolicoeur and Mosimann (1960) that the first principal component, which has maximum variation, may be taken as size factor provided all the coefficients are positive and other principal components with positive and negative coefficients as shape factors. A justification for such an interpretation of principal components is as follows.

Consider the i^{th} element X_i of $\underline{X}$ and the j^{th} principal component $\underline{P}_j'\underline{X}$ of $\underline{X}$. The regression of X_i on $\underline{P}_j\underline{X}$ is simply P_{ji}, the i^{th} element in $\underline{P}_j$. Now, a unit increase in $\underline{P}_j'\underline{X}$ produces on the average an increase of P_{ji} in X_i. If all the coefficients $P_{j1},\ldots,P_{jp}$ are positive, a unit increase in $\underline{P}_j'\underline{X}$ increases the value of each of the measurements, in which case $\underline{P}_j'\underline{X}$ may be called a size factor. If some of the coefficients are positive and others negative, then an increase in $\underline{P}_j'\underline{X}$, increases the values of some of the measurements and decreases the values of the others, in which case $\underline{P}_j'\underline{X}$ may be called a shape factor. Penrose's size and shape functions have similar properties. In either case, there seems to be no suitable interpretation of the relative magnitudes of changes in the different measurements or of the distribution of positive and negative coefficients in the case of a shape factor. There seem to be too many arbitrary elements involved as the principal components are not explicitly derived to represent size and shape factors in any well defined manner.

We shall first generalize the problem by attempting to measure size and shape variation with respect to a given set of measurements $X_1,\ldots,X_p$ in terms of instrumental variables $Z_1,\ldots,Z_m$, where the latter may include some or all of the former.

Let the dispersion matrix of $(\underline{X},\underline{Z})$, where $X' = (X_1,\ldots,X_p)$ and $Z' = (Z_1,\ldots,Z_m)$, be

$$(12.6) \qquad \begin{pmatrix} \underline{\Sigma} & \underline{\Theta} \\ \underline{\Theta}' & \underline{\Gamma} \end{pmatrix}$$

and consider a linear function of $\underline{Z}$, with unit variance,

$$(12.7) \qquad \underline{B}'\underline{Z} = B_1Z_1 + \cdots + B_mZ_m \; , \; \underline{B}'\underline{\Gamma}\;\underline{B} = 1$$

The regression coefficients of $X_1,\ldots,X_p$ on $\underline{B}'\underline{Z}$ are the components of the vector $\underline{\Theta}\ \underline{B}$. Let us specify the ratios of the regression coefficients, taking into account the signs, as the elements of a vector $\underline{R}$, so that

$$\underline{\Theta}\ \underline{B} = \rho\underline{R} \tag{12.8}$$

where ρ is a constant. If all the components of $\underline{R}$ are chosen as positive, then any solution for $\underline{B}$ of the equation (12.8) provides a size function. If some components of $\underline{R}$ are positive and others negative, then a solution for $\underline{B}$ provides a shape function. In the case when $\underline{\Theta}$ is a square matrix of rank m, there is a unique $\underline{B}$ and ρ for a given $\underline{R}$, satisfying the equations

$$\underline{\Theta}\ \underline{B} = \rho\underline{R}\ ,\ \underline{B}'\underline{\Gamma}\ \underline{B} = 1\ . \tag{12.9}$$

Otherwise there may be a multiplicity of solutions. Then $\underline{B}$ may be chosen to maximize ρ.

Introducing Lagrangian multipliers $\underline{C}_1$ (a p-dimensional vector) and c_2 (a constant), the function to be differentiated is

$$\rho + \underline{C}_1'(\underline{\Theta}\ \underline{B} - \rho\underline{R}) + c_2(\underline{B}'\underline{\Gamma}\ \underline{B} - 1). \tag{12.10}$$

Differentiating with respect to ρ, $\underline{B}$, $\underline{C}_1$ and c_2 we obtain the equations

$$\left\{\begin{aligned} \underline{\Theta}'\ \underline{C}_1 + c_2\,\underline{\Gamma}\ \underline{B} &= 0 \\ \underline{\Theta}\ \underline{B} &= \rho\underline{R} \\ \underline{R}'\underline{C}_1 &= 1 \\ \underline{B}'\underline{\Gamma}\ \underline{B} &= 1\ . \end{aligned}\right. \tag{12.11}$$

The first two equations can be written as

$$(12.12) \qquad \begin{cases} \underline{\Theta}'\underline{C} + \underline{\Gamma}\,\rho^{-1}\underline{B} = 0 \\ \underline{\Theta}\,\rho^{-1}\underline{B} = \underline{R} \end{cases}$$

where $\underline{C} = (c_2\rho)^{-1}\underline{C}_1$. Solving the equations (12.12) we have a solution for $\rho^{-1}\underline{B}$, i.e., the size or shape function is determined apart from a constant multiplier. If necessary, the function can be standardized by reducing its variance to unity. To obtain an explicit representation for $\rho^{-1}\underline{B}$, let

$$(12.13) \qquad \begin{pmatrix} \underline{\Theta}' & \underline{\Gamma} \\ \underline{0} & \underline{\Theta} \end{pmatrix}^{-1} = \begin{pmatrix} \underline{A} & \underline{E} \\ \underline{E} & \underline{D} \end{pmatrix}.$$

Then $\rho^{-1}\underline{B} = \underline{D}\underline{R}$. The size and shape functions so determined have greater flexibility, although there will still be some arbitrariness in the choice of the vector of ratios, $\underline{R}$, when it cannot be specified by other considerations.

For instance, anthropologists use the ratio of head breadth to head length, called the cephalic index, to measure the shape of the head. The present approach suggests that the shape function can be built out of a number of length and breadth measurements on the head, by choosing $\underline{R}$ such that the ratios for length measurements have a positive sign and those for breadth measurements have a negative sign. The shape function so determined has the property that an increase in its value increases the lengths and decreases the breadths. Further, other measurements on the body such as stature, chest girth etc., can be brought in to obtain a better measure of head shape. Examples of such functions are given in Rao (1961, 1962a). It has also been found that the choice of absolute values of the elements of $\underline{R}$ in proportion to the standard deviations $\sigma_1, \ldots, \sigma_p$ of the measurements $X_1, \ldots, X_p$ leads to reasonable results.

The size and shape functions determined by the above method may not be uncorrelated as in the case of the principal components, although this can be achieved by choosing the vectors $\underline{R}$ for size and shape functions suitably. But lack of correlation is not an important property if size and shape variations are examined individually. However, if size and shape functions are used jointly in any study, their correlation should be taken into account in the statistical analysis. For instance, if a chart representing the configuration of a set of groups with respect to mean size and shape is desired, one could use 'size' and 'shape corrected for size' (by regression) and represent them on orthogonal axes.

13. Principal component and factor analyses.

Some authors do not make a distinction between these two analyses and this has, no doubt, caused some amount of confusion. It is, therefore, of interest to examine the differences in the nature of information provided by these two techniques. It is shown that they provide distinct answers to two different enquiries concerning a hypothetical structure of the elements of a vector random variable.

Let us suppose that there exist hypothetical uncorrelated factors (variables), $F_1, F_2, \ldots,$ presumably infinite in number, such that

$$\left\{\begin{array}{l} X_1 = a_{11}F_1 + a_{12}F_2 + \cdots \\ \cdot \quad\quad \cdot \quad\quad \cdot \quad \cdot \quad \cdots \\ X_p = a_{p1}F_1 + a_{p2}F_2 + \cdots \end{array}\right. \tag{13.1}$$

or in matrix notation

$$\underline{X} = \underline{A}\,\underline{F}\,. \tag{13.2}$$

It may be seen that the factor structure with common and specific factors, assumed in psychological work is a special (and perhaps an unrealistic!) case of (13.1). The representation (13.1) is not unique for by an orthogonal transformation on the hypothetical variable $\underline{F}$, the structural equation (13.2) can be written as

$$\underline{X} = \underline{B}\,\underline{G} \tag{13.3}$$

where $\underline{G}$ represents a new (transformed) factor variable.

The problem we wish to investigate is one of approximating $X_1, \ldots, X_p$ by linear functions of as few factors, which may be called dominant, as possible. An approximation to $\underline{X}$ in terms of q factors, to be denoted by $\underline{X}^{(q)}$ is obtained by considering a representation of $\underline{X}$, such as (13.1) and truncating the right hand side at the q^{th} term. In matrix notation

$$\underline{X}^{(q)} = \underline{A}_q \underline{F}^{(q)} \tag{13.4}$$

where $\underline{A}_q$ is the matrix of the first q columns of $\underline{A}$ and $\underline{F}^{(q)}$ is the vector of the variables $F_1, \ldots, F_q$. The different choices of $\underline{A}_q$ are obtained by making orthogonal transformations on $\underline{F}$ (the entire set of hypothetical factors), leading to different representations of $\underline{X}$ and applying the truncation procedure.

We have not yet specified the nature of the approximation we are seeking. It is seen that the dispersion matrix of $\underline{X}^{(q)}$ is $\underline{A}_q \underline{A}_q'$ (assuming without loss of generality that all F_i have unit variance), while the dispersion matrix of $\underline{X}$ is $\underline{\Sigma}$. Let us suppose that the approximation aims at the closeness of $\underline{\Sigma}$ and $\underline{A}_q \underline{A}_q'$. In such a case, we may determine $\underline{A}_q$ by minimizing the Euclidean norm of the residual dispersion matrix $\|\underline{\Sigma}-\underline{A}_q\underline{A}'_q\|$. Applying the result (2.6), the answer is

$$\underline{A}_q = (\sqrt{\lambda_1}\,\underline{P}_1 \;\; \sqrt{\lambda_2}\,\underline{P}_2 \cdots \sqrt{\lambda_q}\,\underline{P}_q) \tag{13.5}$$

i.e., the i^{th} column of $\underline{A}_q$ is $\sqrt{\lambda_i}\,\underline{P}_i$ where λ_i is the i^{th} eigen value and $\underline{P}_i$ is the i^{th} eigen vector of $\underline{\Sigma}$. Then it may be shown that the best estimates of the factors are

$$F_i^q = \lambda_i^{-1/2}\underline{P}'_i\underline{X}, \quad i=1,\dots,q \tag{13.6}$$

where $\underline{P}'_i\underline{X}$ is the i^{th} principal component of $\underline{X}$. We are thus led to principal component analysis as a definite answer to the problem posed.

Using estimated F_i^q, the q^{th} approximation to $\underline{X}$ as in (13.6), the residual is

$$\underline{X}^q-\underline{X} = (\underline{P}_{q+1}\underline{P}'_{q+1} + \dots + \underline{P}_p\underline{P}'_p)\underline{X}\,. \tag{13.7}$$

The dispersion matrix of the residual is

$$\begin{aligned}&(\underline{P}_{q+1}\underline{P}'_{q+1} + \dots + \underline{P}_p\underline{P}'_p)\,\underline{\Sigma}\,(\underline{P}_{q+1}\underline{P}'_{q+1} + \dots + \underline{P}_p\underline{P}_p)' \\ &= \lambda_{q+1}P_{q+1}P'_{q+1} + \dots + \lambda_p P_p P'_p\end{aligned} \tag{13.8}$$

which is small if $\lambda_{q+1}, \dots, \lambda_p$ are small. In any problem we can examine the actual residuals (13.7) in judging the adequacy of a given order of approximation.

We have seen how the criterion of minimizing $\|\underline{\Sigma}-\underline{A}_q\underline{A}'_q\|$ led to the principal component analysis. We shall show that a different criterion leads to factor analysis.

The amount of correlation between X_i, X_j explained by the q factors $F_1,\dots,F_q$ is

$$\mathrm{cov}(X_i^q,\ X_j^q) \div \sqrt{V(X_i)V(X_j)}$$

$$= (a_{i1}a_{j1} + \cdots + a_{iq}a_{jq})/\sigma_i\sigma_j$$

$$= b_{i1}b_{j1} + \cdots + b_{iq}b_{jq}$$

where σ_i is the standard deviation of X_i and $b_{ik} = a_{ik}/\sigma_i$. The matrix of all possible correlations explained by the q factors is then the off diagonal elements of $\underline{B}_q\underline{B}_q'$ where $\underline{B}_q$ is the matrix with its i^{th} row as $(b_{i1},\ldots,b_{iq})$. Let $\underline{\Lambda}$ be the actual correlation matrix of the variables $X_1,\ldots,X_p$. We wish to choose $\underline{B}_q$ such that the off diagonal elements of $\underline{\Lambda}-\underline{B}_q\underline{B}_q$ are as small as possible. A single measure of difference is the sum of squares of the off diagonal elements of $\underline{\Lambda}-\underline{B}_q\underline{B}_q'$ and the mathematical problem is one of minimizing this measure subject to the condition that the diagonal entries of $\underline{\Lambda}-\underline{B}_q\underline{B}_q$ are non-negative. The factor structure provided by the optimum $\underline{B}_q$ is known as factor analysis; the attempt in such a case is to explain the correlation among the measurements rather than the variances of the individual measurements. The problem of an optimum choice of $\underline{B}_q$ is not easy to solve, and no definite solution can be obtained. An iterative method is suggested by Rao (1955) but the convergence of the procedure has not been adequately studied. For further literature see Whittle (1953), and the references in Maxwell and Lawley (1963).

14. An application to a problem in multidimensional scaling.

The data consist of $n(n-1)/2$ independent determinations (possibly subject to errors) of distances between n stimuli and the problem is to determine the smallest Euclidean space in which the stimuli can be represented as points (Torgerson, 1958). Let $\underline{X}_1, \ldots, \underline{X}_n$ be points representing the n stimuli in a Euclidean space, and $\overline{X}$ be the center of gravity of the points. Then

$$
\begin{aligned}
&(\underline{X}_i-\overline{X})'(\underline{X}_j-\overline{X}) \\
&\quad = n^{-2}[(\underline{X}_i-\underline{X}_1)+\cdots+(\underline{X}_i-\underline{X}_n)]'[(\underline{X}_j-\underline{X}_1)+\cdots+(\underline{X}_j-\underline{X}_n)] \\
(14.1) \qquad &\quad = n^{-2}\sum_{k,}\sum_{m}[(\underline{X}_j-\underline{X}_k)'(\underline{X}_j-\underline{X}_k)+(\underline{X}_i-\underline{X}_m)'(\underline{X}_i-\underline{X}_m) \\
&\qquad\qquad - (\underline{X}_i-\underline{X}_j)'(\underline{X}_i-\underline{X}_j)-(\underline{X}_k-\underline{X}_m)'(\underline{X}_k-\underline{X}_m)] \\
&\quad = \frac{1}{2n^2}\sum_{k,}\sum_{m}(\delta_{jk}+\delta_{im}-\delta_{ij}-\delta_{mk})
\end{aligned}
$$

where δ_{rs} denotes the square of the true distance between the points $\underline{X}_r$ and $\underline{X}_s$. Using the notation $\sum_m \delta_{im} = \delta_{i.}$ and $\sum_{k,}\sum_m \delta_{km} = \delta_{..}$ the expression (14.1) can be written

$$
(14.2) \qquad \tfrac{1}{2}\left(-\delta_{ij} + \frac{\delta_{i.}+\delta_{.j}}{n} - \frac{\delta_{..}}{n^2}\right).
$$

Thus, given the mutual distances we can express the configuration of the points by the matrix $\underset{\sim}{X}_d' \underset{\sim}{X}_d$ whose (i,j) entry is given by (14.2), where $\underset{\sim}{X}_d$ is as defined in (3.3). The matrix $\underset{\sim}{X}_d' \underset{\sim}{X}_d$, as introduced in section 4, represents the lengths of the lines joining the points to the center of gravity and the angles between them. The best representation of the points in a Euclidean space of q dimensions is provided by the $q \times m$ matrix whose i^{th} row is $\sqrt{\lambda_i}\, Q_i$ where λ_i and Q_i are the i^{th} eigen value and vector of $\underset{\sim}{X}_d' \underset{\sim}{X}_d$.

However, such an analysis is not directly applicable when the exact distances δ_{ij} are not known, but only estimates d_{ij} of δ_{ij} are available. Substituting d_{ij} for δ_{ij} in (14.2) we have an estimate of $\underset{\sim}{X}_d' \underset{\sim}{X}_d$ which we may denote by $B = (b_{ij})$. The matrix $\underline{B}$ need not be non-negative definite. We may still compute the eigen values and vectors of $\underline{B}$. If $\hat{\lambda}_i$ and $\hat{Q}_i$ denote the i^{th} eigen value and vector of $\underline{B}$, an estimate of the best q dimensional representation is given by the $q \times m$ matrix whose i^{th} row is $\sqrt{\hat{\lambda}_i}\,\hat{Q}_i$, provided $\hat{\lambda}_1, \ldots, \hat{\lambda}_q$ are positive.

One drawback of the solution derived from B is that the estimated points may not have the origin as the center of gravity. This may be secured by determining the eigen vectors of B with the restriction that the sum of the elements of each eigen vector is zero. Let U represent a column vector of n unities. Then it can be shown, using result (iii) of section 2, that the eigen vectors of the matrix

$$(\underline{I} - \frac{\underline{U}\,\underline{U}'}{n})\underline{B} \tag{14.3}$$

satisfy the required conditions. Let $\mu_1, \mu_2, \ldots$ be the eigen values and $\underline{B}_1, \underline{B}_2, \ldots$ the corresponding eigen vectors of the matrix (14.3). Then the rows of the estimated matrix of points (columns representing the points) in a q dimentionsl space are $\sqrt{\mu_1}\,B_1', \ldots, \sqrt{\mu_q}\,B_q'$.

15. Estimation of structural relationship, hetergeneity of data, etc.

A recent application by Wernimont (1963) demonstrates the use of principal component analysis in detecting hetergeneity of data. In his study there are 9 spectrophotometers and each spectrometer was used to obtain a series of six absorbance curves (as a discrete set measurements at 20 different wave lengths) of solutions at three different concentrations (30, 60 and 90) and on two different days. The data consisted of 54 sets of 20 measurements and in the notation of the present paper $n = 54$ and $p = 20$. According to Beer's law, if the same wave lengths have been used for all the spectrometers, the 54 points in the 20 dimensional space should ideally be on a line through the origin when there are no errors in the absorbance readings. The direction cosines of such a line would provide the constants of Beer's law for a given set of wavelengths. If, indeed, the wavelength settings differed from one instrument to another the configuration of points would not be confined to a straight line. The object of investigation was to examine whether the instruments had systematic errors in the wavelength setting. If some of the instruments did not allow a proper setting of the desired wave lengths, the data would be heterogeneous; this could be detected by asking the question whether the observed points could be regarded as clustering round a straight line.

This is exactly what the principal components can detect provided the errors of absorbance measurements at each wavelength are independent and have nearly the same distribution.

According to Beer's law, the vector $\underline{X}$ of measurements at 20 different wavelenghts can be written as

$$\underline{X} = \underline{\beta} c + \underline{\epsilon} \tag{15.1}$$

where $\underline{\beta}$ depends on the wavelengths, c is the concentration of the solution used and $\underline{\epsilon}$ represents the errors of absorbance readings. Thus for different values of c, the true locus of the points $\underline{X}$ is a straight line through the origin, with direction cosines proportional to $\underline{\beta}$. But if $(\underline{\beta} + \underline{\xi}_{ij})$ is the vector appropriate to the i^{th} spectrophotometer on the j^{th} day, where ξ_{ij} represents the systematic error, then for the measurements with the i^{th} spectrophotometer on the j^{th} day at concentration c

$$(15.2) \qquad \underline{X} = (\underline{\beta} + \underline{\xi}_{ij})c + \underline{\epsilon}$$

so that Beer's law holds for a given instrument on a given day. In such a case we should expect the 54 points to cluster round a maximum of 18 different lines corresponding to the nine instruments and two days. The 18 lines may lie in a space of as many as 18 dimensions depending on the complexity of the systematic errors. Let us examine this problem by the principal component analysis by pooling the data from all the instruments. The 54 vectors provide a 20 × 20 dispersion matrix. [Actually in a problem with the model such as (15.2), the principal component analysis could be carried out on the uncorrected dispersion matrix. The computations referred to are on the corrected dispersion matrix as reported by Wernimont.] The first two eigen values of this matrix have been found to be (0.119590×10^7) and (0.356424×10^3) which explain 99.94 and 0.03 percent of total variation respectively. The rest of the eigen values are much smaller indicating that the configuration of the points can be examined by one or two dimensional representations.

To find the coordinates of the projected points on the best two dimensional plane we have to determine the first two eigen vectors $\underline{P}_1$ and $\underline{P}_2$ and compute the coordinates

$$(15.3) \qquad \underline{P}_1'X \ , \ \underline{P}_2'X$$

corresponding to each original point X. Transferring the original to $P_1'\bar{X}$, $P_2'\bar{X}$, the coordinates are

$$\underline{P}_1'(\underline{X}-\underline{\bar{X}}),\ \underline{P}_2'(\underline{X}-\underline{\bar{X}})\ . \tag{15.4}$$

For each point, it is also necessary to compute the length of the perpendicular on the best fitting plane to examine whether the representation is adequate with respect to that particular point. The square of the perpendicular from the point X is

$$R^2 = (\underline{X}-\underline{\bar{X}})'(\underline{X}-\underline{\bar{X}})-[\underline{P}_1'(\underline{X}-\underline{\bar{X}})]^2-[\underline{P}_2'(\underline{X}-\underline{\bar{X}})]^2\ . \tag{15.5}$$

The following table gives the values of (15.4) scaled by the reciprocal of the square root of the eigen values [which is not necessary for our purpose, but adopted from the computations provided by Wernimont] and R^2 for two sets of six absorbance curves of S.P. Meters 1 and 7. Similar values are available for all the 54 absorbance curves.

No.	S.P.M.	Concentration	Day	$\underline{P}_1'(\underline{X}-\underline{\bar{X}})/\sqrt{\lambda_1}$	$\underline{P}_2'(\underline{X}-\underline{\bar{X}})/\sqrt{\lambda_2}$	R^2
1.1	1	30	1	-1.220	0.614	54.17
1.2		60	1	-0.016	1.419	254.51
1.3		90	1	1.173	1.875	443.93
1.4		30	2	-1.200	0.574	122.15
1.5		60	2	0.001	1.518	372.83
1.6		90	2	1.200	1.877	334.46
7.1	7	30	1	-1.235	0.590	715.20
7.2		60	1	-0.065	0.140	1166.08
7.3		90	1	1.148	-0.266	714.84
7.4		30	2	-1.210	0.649	405.33
7.5		60	2	-0.061	0.261	610.58
7.6		90	2	1.134	0.074	655.72

Using the coordinates $(\underline{P}_1'(\underline{X}-\underline{\bar{X}})/\sqrt{\lambda_1},\ \underline{P}_2'(\underline{X}-\underline{\bar{X}})/\sqrt{\lambda_2})$, the 54 points are plotted on two charts (to avoid over crowding). In the charts (A·1, A·2, A·3) and (A·4, A·5, A·6) represent the absorbance curves at three levels of concentration obtained with S_A (the spectrophotometer A) on the first and second days respectively. The following conclusions emerge. (a) Generally, the points (A·1, A·2, A·3) are close to a line and so also the points (A·4, A·5, A·6) indicating the validity of Beer's law. (b) The points (A·1, A·2, A·3) and (A·4, A·5, A·6) lie on the same line except for S_6 and S_9, indicating errors specific to the instruments. (c) The lines for different instruments are differently oriented indicating systematic errors.

16. Determination of clusters of points.

Let $\underline{X}_1, \ldots, \underline{X}_n$ be points in a p-dimensional space. In some problems it is of interest to examine whether the n points can be classified into groups or clusters such that the points within a cluster are close together, while the clusters themselves are far apart. Thus if the p coordinates of each point represent the production of p agricultural commodities in a region and different points represent different regions we may like to group together regions which are similar with respect to over all agricultural production (Kendall, 1939). Or if the p coordinates of each point represent the mean anthropological characters of a population and different points represent different populations we may like to examine whether the populations can be grouped into distinct clusters on the basis of similarity of the characters

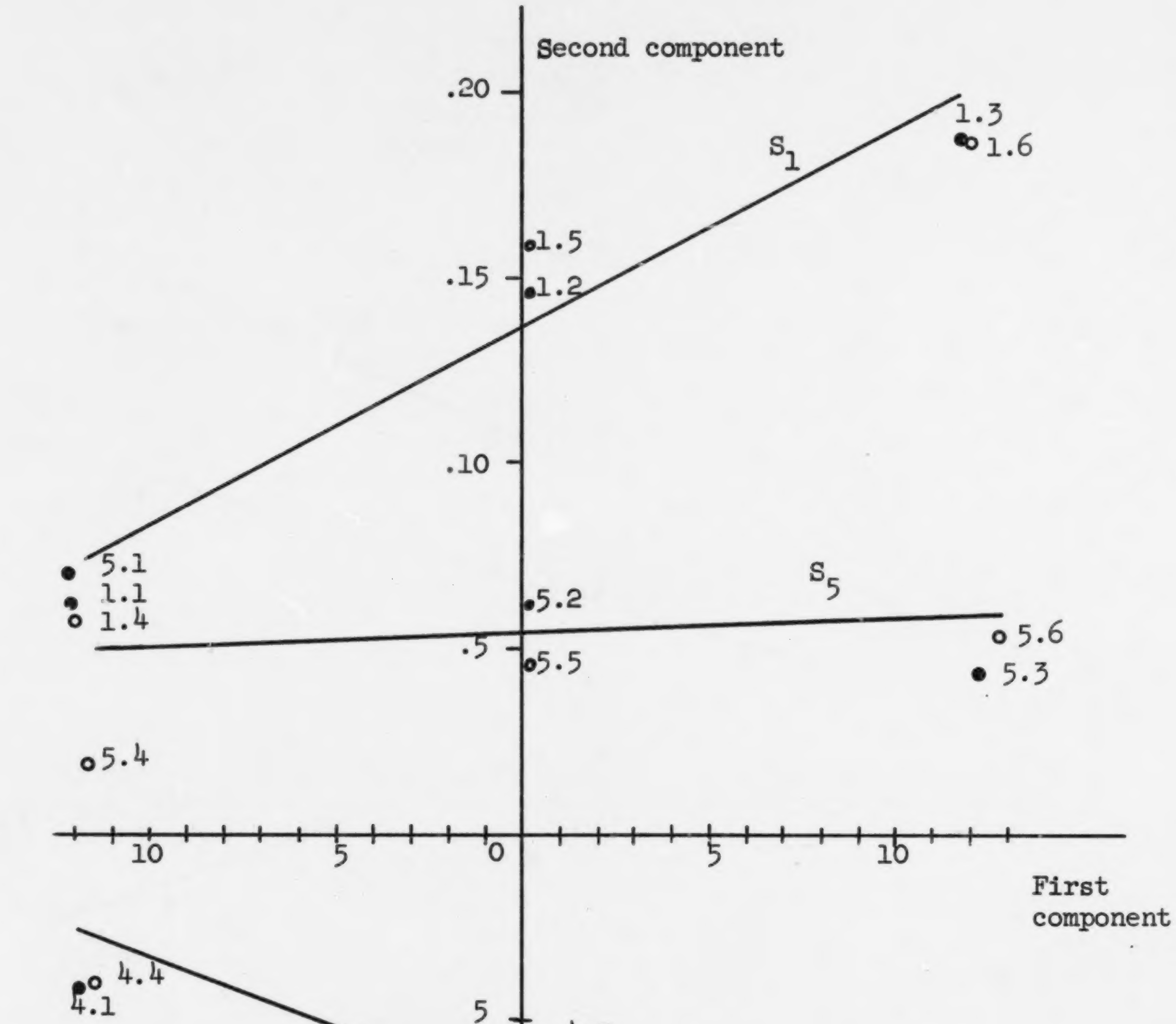


Chart 1. Configuration of the points with respect to the first two Principal Components.

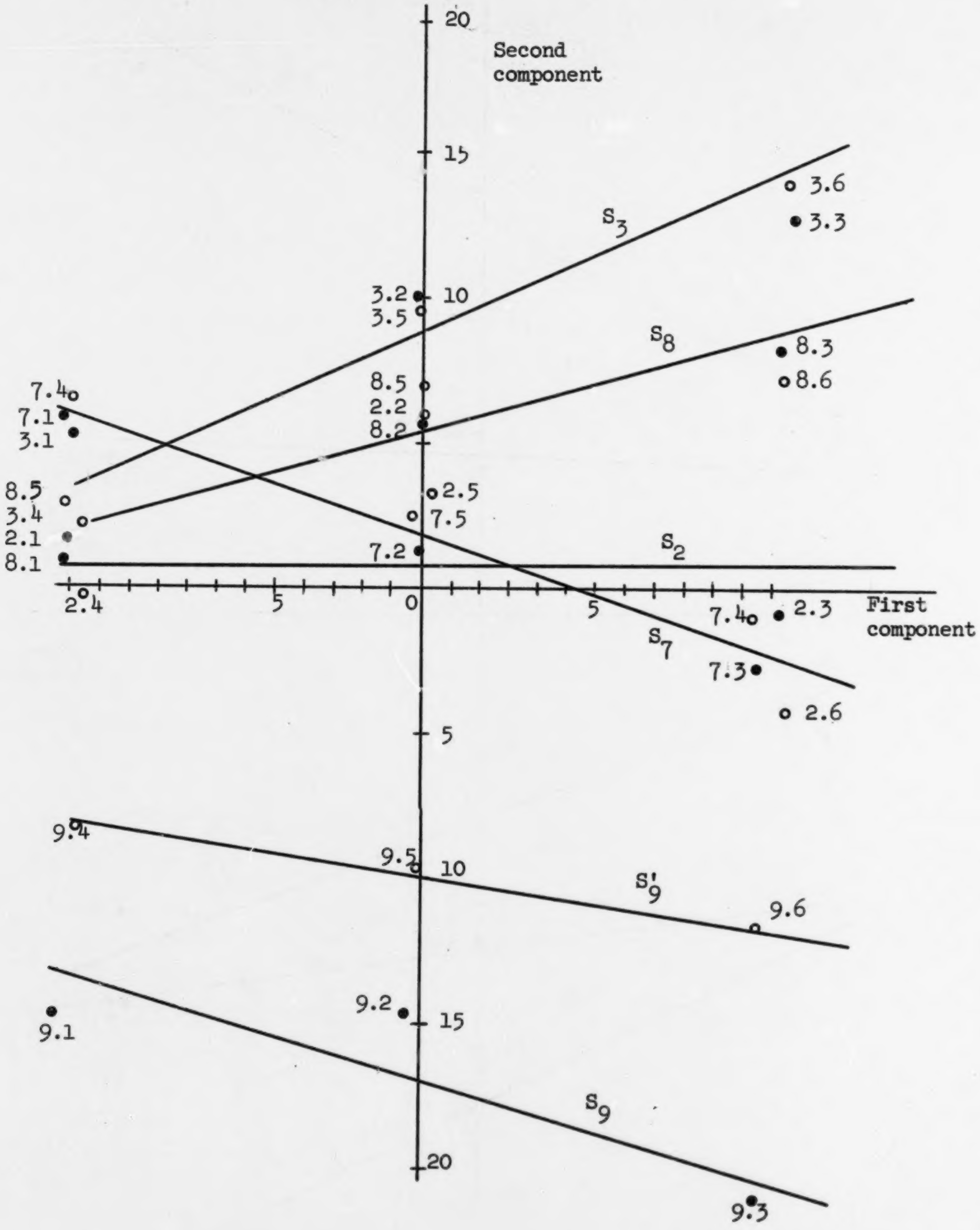


Chart 2. Configuration of the points with respect to the first two Principal Components.

(Mahalanobis, 1936; Mahalanobis, Majumdar and Rao, 1949) etc.). In all such situations we need to have a measure of distance between two points. One may choose as the distance between $\underline{X}_i$ and $\underline{X}_j$

$$d_{ij} = (\underline{X}_i-\underline{X}_j)'(\underline{X}_i-\underline{X}_j) \tag{16.1}$$

representing the points in a Euclidean space with orthogonal axes or

$$d_{ij} = (\underline{X}_i-\underline{X}_j)\underline{\Lambda}^{-1}(\underline{X}_i-\underline{X}_j) \tag{16.2}$$

representing the points in a Euclidean space with oblique axes. When $\underline{\Lambda}$ is the covariance matrix of the characters within a population (or region) the distance (16.2) is known as Mahalanobis distance.

Whatever distance function is chosen, we have a $n \times n$ matrix $D = (d_{ij})$ of all possible distances. We, then, look for clusters of points, by fixing a few pairs of points which are close together, but the pairs themselves being more distant from each other, and building clusters around them by adjoining neighboring points. In practice such a program can be successfully carried out although considerable difficulties are involved (see Mahalanobis, Majumdar and Rao, 1949; Majumdar and Rao, 1958).

When $p = 1$, we have one dimensional ordering of points and no visual representation of points on a line is necessary to determine the clusters. When $p = 2$, the determination of clusters is also simple, since we can represent the points on a two dimensional chart (using orthogonal or oblique axes) and mark out the subsets of points close together. For $p = 3$, we may use a three dimensional model. But no such visual aid is available for larger values of p.

We may, then raise the question as to whether the points in a p-dimensional space can be reasonably represented in 2 or 3 dimensional spaces without distorting the configuration as specified by the mutual distances. We have seen in sections 3 and 4, that such a representation is secured through the principal components. Let $\underline{P}_1, \underline{P}_2, \underline{P}_3$ be the first three eigen vectors of $\underline{\Sigma}$ (or $\underline{\Sigma}$ with respect to $\underline{\Lambda}$ in the oblique case) giving the coordinates $\underline{P}_1'\underline{X}_i$, $\underline{P}_2'\underline{X}_i$, $\underline{P}_3'\underline{X}_i$ for the representation of $\underline{X}_i$ in the best three dimensional space. We may choose, the first, or the first two or all the three coordinates to obtain a visual representation of the points and then determine the clusters of points.

The success of the method obviously depends on how well the configuration of the points is preserved in the reduced space. If $\lambda_1, \ldots, \lambda_p$ are all the eigen values an overall measure of the adequacy of the best q dimensional representation as introduced in section 3 is $(\lambda_1 + \cdots + \lambda_q)/(\lambda_1 + \cdots + \lambda_p)$. We may examine the values of this measure for $q = 1,2,3$ and decide on a suitable value of q. But such a decision may be misleading in individual cases. So it is necessary to obtain a tabulation of all the distances in the original and reduced space for a preliminary comparison. The distance between the ith and jth points in the reduced space of q dimensions is computed by the formula

$$d_{ij}^q = (\underline{P}_1'\underline{X}_i - \underline{P}_1'\underline{X}_j)^2 + \cdots + (\underline{P}_q'\underline{X}_i - \underline{P}_q'\underline{X}_j)^2 \tag{16.3}$$

whether the original points are in an orthogonal space or not. The difference $d_{ij} - d_{ij}^q$ is always positive and measures the efficiency of reduction in the dimensions of the space with respect to the points i and j. If the differences $(d_{ij} - d_{ij}^q)$ are uniformly small, i.e., for all i and j, we are on safe grounds in determining the clusters on the basis of the first q principal components. If not we can still use the q principal components for the points where the distances are not distorted. The affinities of the points whose distances with the others are distorted may then be determined by considering their actual distances between themselves and with the others.

17. Tests of significance for principal components.

In the previous sections no reference has been made to tests of significance in principal component analysis. The reason is that the problems have not posed in terms of testing of well defined hypotheses but in terms of estimating some features of the measurements on the individuals of a population. It is assumed that in any realistic problem, reduction in the number of measurements (by omission of some or by replacing the original measurements by a smaller number of linear functions) entails some loss of information. If so we can only consider null hypotheses which specify the amount of information lost. For instance, a null hypothesis may specify that the ratio of the sum of the last $p-q$ eigen values to the sum of all eigen values of a hypothetical dispersion matrix Σ is π. How can we test such a hypothesis when we have an estimate $\hat{\Sigma}$ of Σ. The relevant statistic appears to be

$$(\hat{\lambda}_{q+1} + \cdots + \hat{\lambda}_p)/(\hat{\lambda}_1 + \cdots + \hat{\lambda}_p) \tag{17.1}$$

where $\hat{\lambda}_1, \ldots, \hat{\lambda}_p$ are the eigen values of $\hat{\Sigma}$. It is extremely difficult to derive the distribution of (17.1) even under the assumption that $\hat{\Sigma}$ has a Wishart distribution.

Another null hypothesis of interest is the equality of the last $p-q$ true eigen values. An appropriate statistic to test such a hypothesis is the ratio of the geometric mean to the arighmetic mean of the estimated roots $\hat{\lambda}_{q+1},\dots,\hat{\lambda}_p$, as shown by (Bartlett, 1950, 1951a, 1951b; Rao 1955). The large sample distribution of the logarithm of the ratio suitably standardized is chi-square when $\hat{\Sigma}$ has a Wishart distribution. However, the tests based on the roots $\hat{\lambda}_1,\dots,\hat{\lambda}_p$ are likely to be sensitive to departure from Wishartness of the distribution of $\hat{\Sigma}$ and no appropriate robust test criteria have been found.

18. Use of principal components in testing differences in mean values between groups.

It is well known that when multiple measurements are available on samples of individuals from different populations, the total sum of products (S.P.) matrix can be analyzed as due to 'within' and 'between' groups. The analysis of dispersion (which is a generalization of analysis of variance) is indicated as follows (Rao, 1952).

	D.F.	S.P. matrix
Between	k-1	B
Within	n	W
	n+k-1	T

Differences between the mean values of k populations and the configuration of the mean values when differences exist are examined by the roots of the determinental equation

$$|B-\theta T| = o \quad . \tag{18.1}$$

For applications the reader is referred to (Bartlett, 1948, Fisher, 1939; Rao, 1948, 1952; Williams, 1959, etc.).

Recently Roy (1958) proposed stepdown procedures whereby differences are examined in one observed variable first, and then in another observed variable eliminating the regression due to the first, and so on using each time a univariate procedure. There is, however, the problem of fixing the order in which the variables are considered. It is suggested that instead of the original variables, the principal components

$$\underline{P}_1'\underline{X}, \ldots, \underline{P}_q'\underline{X} \tag{18.2}$$

where $\underline{P}_1, \ldots, \underline{P}_q$ are the first q eigen vectors of the total S.P. matrix T may be considered in the given order. The performance of such a procedure (Dempster, 1963) is not fully studied except in a very special case arising in the comparison of growth curves (Rao, 1958). Although the method suggested is in line with the general philosophy of the principal component analysis it is clear that the first few principal components (18.2) are not designed to summarize as much information as possible on the differences between the mean values of the groups. In actual practice it may turn out that some of the principal components with smaller eigen values are better discriminators between the populations with respect to the differences in means than the first few principal components. Hence the principal components ordered by the eigen values of T may not be necessarily ordered with respect to their discrimination power. This is so only in special cases such as those considered by Rao (1958). However, the general draw back of the step-down procedures is that they do not enable us to study the configuration of the mean values of the different populations, which is more important than merely establishing differences in the mean values in practical problems.

REFERENCES

[1] Bartlett, M.S. (1947), "Multivariate Analysis," *J. Roy. Statist. Soc.* (supple.) **9**, 176-197
(1950) "Tests of Significance in Factor Analysis," *Brit. J. Psych.* (Stat. Sec.), **3**, 77-85
(1951) "The Effect of Standardization of a X^2-Approximation in Factor Analysis," *Biometrika*, **38**, 337-344
(1951) "A Further Note on Tests of Significance in Factor Analysis," *Brit. J. Psych*, (Stat. Sec.), **4**, 1-2

[2] Dempster, A.P (1963), "Multivariate Theory for General Stepwise Methods," *Ann. Math. Statist.*, **34**, 873-883

[3] Fisher, R.A. (1939), "The Sampling Distribution of Some Statistics Obtained From Non-Linear Equations," *Ann. Eugen.*, **9**, 238-249

[4] Frisch, R. (1929), "Correlation and Scatter in Statistical Variables," *Nordic Stat. J.*, **8**, 36-102

[5] Hotelling, H. (1933), "Analysis of a Complex Statistical Variable into Principal Components," *J. Educ. Psych.*, **26**, 417-441, 498-520
(1935) "The Most predictable Criterion," *J. Educ. Psych.*, **26**, 139-142
(1936) "Simplified Calculation of Principal Components," *Psychometrika*, **1**, 27-35
(1936) "Relations Between Two Sets of Variates," *Biometrika*, **28**, 321-377

[6] Jolicoeur, P. and Mosimann, J.E. (1960), "Size and Shape Variation in the Painted Turtle, A Principal Component Analysis," *Growth*, **24**, 339-354

[7] Kendall, M.G. (1939), "The Geographical Distribution of Crop Productivity in England," *J. Roy. Statist. Soc.*, **102**, 21-

[8] Mahalanobis, P.C. (1936), "On the Generalized Distance in Statistics," Proc. Nat. Inst. Sci. India, 12, 49-55

[9] Mahalanobis, P.C., Majundar, D.N. and Rao, C.R. (1949), "Anthropometric Survey of the United Provinces, 1941. A Statistical Study," Sankhya, 9, 90-327

[10] Majundar, D.N. and Rao, C.R. (1958), "Bengal Anthropometric Survey, 1945. A Statistical Study," Sankhya, 19, 201-408

[11] Mosimann, J.E. (1958), "An Analysis of Allometry in Chelonian Shell," Rev. Can. Biol., 17, 137-228

[12] Pearson, K. (1901), "On Lines and Planes of Closest Fit to Systems of Points in Space," Phil. Mag., 2(sixth series), 559-572

[13] Penrose, L.S. (1947), "Some Notes on Discrimination," Ann. Eugen., 13, 228-237

[14] Rao, C.R. (1948), "Tests of Significance in Multivariate Analysis," Biometrika, 35, 58-79
(1948) "On a Formula for the Prediction of Cranial Capacity," Biometrics, 4, 247-253 (with D.C. Shaw)
(1952) Advanced Statistical Methods in Biometric Research, John Wiley and Sons
(1955) "Estimation and Tests of Significance in Factor Analysis," Psychometrika, 20, 93-111
(1958) "Some Statistical Methods for Comparison of Growth Curves," Biometrics, 14, 1-17
(1960) "Multivariate Analysis: An Indispensable Aid in Applied Research," Sankhya, 22, 317-338
(1961) "Some Observations on Multivariate Statistical Methods in Anthropological Research," Bull. Inst. Statist. Inst., 38, 99-109
(1962) "Use of Discriminant and Allied Function in Multivariate Analysis," Sankhya, 22, 317-338

Rao, C.R. (continued)

(1962) "Problems of Selection with Restrictions," *J. Roy. Statist. Soc.*, (series B), 24, 401-405

(1964) "Problems of Selection Involving Programming Techniques," IBM Symposium on Statistics

[15] Roy, J. (1958), "Step-Down Procedure in Multivariate Analysis," *Ann. Math. Statist.*, 29, 1177-1187

[16] Simonds, J.L. (1963) "Applications of Characteristic Victor Analysis to Photographic and Optical Response Data," *J. Optical Society of America*, 53, 968-974

[17] Stone, R. (1947), "An Interdependence of Blocks of Transactions," *J. Roy. Stat. Soc. Supple.*, 9, 1-32

[18] Torgerson, W.S. (1958), *Theory and Methods of Scaling*, John Wiley and Sons

[19] Wernimont, G. (1963), "Use of the Computer to Compare Spectrophotometric Curves," Management Systems Development Department. Eastman Kodak Company

[20] Whittle, P. (1953), "On Principal Components and Least Square Methods of Factor Analysis," *Skand. Aktuarictidskr.* 36, 223-239

[21] Williams, E.G. (1958), "*Regression Analysis*," John Wiley and Sons

TESTING FOR THE EQUALITY OF CORRELATION COEFFICIENTS FOR VARIOUS MULTIVARIATE MODELS

by

Ingram Olkin and Minoru Siotani

1. Introduction.

The genesis of the present investigation stems from the following problem. Samples are drawn from two bivariate normal populations with correlation coefficients ρ_1 and ρ_2, and we wish to test for the equality of ρ_1 and ρ_2. A natural procedure based on large sample theory is to transform the sample correlation coefficients r_1 and r_2 to Fisher's $z = \frac{1}{2}\log[(1+r)/(1-r)]$, see e.g., Walker and Lev [5]. Since $z_1 - z_2$ is approximately normally distributed with mean $\xi_1 - \xi_2$ and variance $(N_1-3)^{-1} + (N_2-3)^{-1}$, where $\xi_i = \frac{1}{2}\log[(1+\rho_i)/(1-\rho_i)]$, we can use $z_1 - z_2$ as a test statistic.

In this context, Walker and Lev [5] refer to data of Durost [2] concerning children's collecting activities in which measurements on mental age and ratings on the child's collections were made for 50 boys and 50 girls between the ages 10 and 14. The question here was whether the relation between mental maturity and quality of collections is higher for boys than for girls.

The above procedure, however, does not take cognizance of the structure of the variances, so that it is reasonable to expect different estimators and procedures depending on the particular underlying model.

The purpose of the present paper is to obtain maximum likelihood estimators (MLE) of the relevant parameters for several different models. It will then be an easy conversion to obtain likelihood ratio tests.

Section 2 deals with preliminaries. In Section 3, maximum likelihood estimators for the following models are obtained.

Model	Population	Parameters
	I	II
A	$\sigma_{11}, \sigma_{22}, \rho_1$	$\tau_{11}, \tau_{22}, \rho_2$
B	$\sigma_{11}, \sigma_{22}, \rho$	$\tau_{11}, \tau_{22}, \rho$
C	$\sigma_{11}, \sigma_{22}, \rho_1$	$\sigma_{11}, \sigma_{22}, \rho_2$
D	$\sigma_{11}, \sigma_{22}, \rho$	$\sigma_{11}, \sigma_{22}, \rho$

Of particular interest is Model B, for this is the structure most frequently encountered. Even in this case, the use of the z- transformation does not in general yield a MLE. However, when $N_1 = N_2$, the use of z is equivalent to the MLE.

In Section 4 we comment on the k-sample problem, and in Section 5 discuss tests of hypotheses.

By an intraclass correlation model, we mean that we have a p-variate normal population with homogeneous variances and homogeneous covariances. Thus the relevant parameters are $\sigma_{11} = \cdots = \sigma_{pp} = \sigma^2$, and $\sigma_{ij} = \sigma^2\rho$ $(i \neq j)$. In Section 6, MLE for the following models are discussed.

Model	Population	Parameters
	I	II
a	σ^2, ρ_1	τ^2, ρ_2
b	σ^2, ρ	τ^σ, ρ

2. Preliminaries.

We first consider a general formulation for the p-variate problem, and then specialize to the case $p=2$. Let the sample covariance matrix S have a Wishart distribution, denoted by $\mathcal{L}(S) = W(\Sigma; p, n)$, with density function

$$(2.1) \qquad p(S) = c(p,n;S)\ |\Sigma|^{-n/2} \exp - \frac{1}{2} \operatorname{tr} \Sigma^{-1} S\ , \qquad S > 0\ ,$$

$$c(p,n;S) = |S|^{(n-p-1)/2} [2^{pn/2}\ \pi^{p(p-1)/4} \prod_{1}^{p} \Gamma(\tfrac{n-i+1}{2})]^{-1}\ .$$

In order to separate the variances from the correlation coefficients, we write

$$\Sigma = D_\sigma\ P\ D_\sigma\ ,$$

where $D_\sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_p)$, $\sigma_i^2 = \sigma_{ii}$, and $P = (\rho_{ij})$, $\rho_{ii} = 1$, is the correlation matrix.

At times it will be easier to maximize with respect to $\lambda_j = 1/\sigma_j$, and we let $D_\lambda \equiv D_\sigma^{-1}$. Also, we may wish to first maximize with respect to the λ's. To accomplish this, we note that

$$\operatorname{tr}\ \Sigma^{-1} S = \operatorname{tr} D_\lambda\ P^{-1}\ D_\lambda\ S = \sum_{i,j} \lambda_i\ \lambda_j\ \rho^{ij}\ s_{ij}\ .$$

This suggests that we define $A = P^{-1} \times S$, i.e., $a_{ij} = \rho^{ij}\ s_{ij}$, so that

$$\operatorname{tr} \Sigma^{-1} S = \lambda A \lambda'\ , \quad \lambda = (\lambda_1, \ldots, \lambda_p)\ .$$

We also note that if P^{-1} and S are positive definite then A is positive definite (Schur's Theorem, e.g., [1]). Thus, to maximize $p(S)$

with respect to λ, we need only consider the terms of (2.1) which involve λ, namely

$$\max_{\lambda} |D_\lambda|^n \exp - \frac{1}{2} \lambda A \lambda' .$$

Setting the derivatives of the logarithm equal to zero yields the matrix equation

$$(2.2) \qquad n(\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_p}) = (\lambda_1, \ldots, \lambda_p) A .$$

From this we immediately obtain

$$(2.3) \qquad np = \lambda A \lambda' .$$

A general solution to (2.2) seems untractable. However, it was shown by Marshall and Olkin [3] in another context that there is a solution in the positive orthant. When $p=2$, we can give an explicit solution. The two equations become

$$n = \lambda_1^2 a_{11} + \lambda_1 \lambda_2 a_{12} ,$$

$$n = \lambda_2^2 a_{22} + \lambda_1 \lambda_2 a_{12} ,$$

so that $\lambda_1^2 a_{11} = \lambda_2^2 a_{22}$. Because $0 < \lambda_1$, $0 < a_{11}$, we have that $\lambda_2 = \lambda_1 \sqrt{a_{11}/a_{22}}$, and

$$(2.4) \qquad \hat{\lambda}_1^2 = \frac{n\sqrt{a_{22}/a_{11}}}{\sqrt{a_{11} a_{22}} + a_{12}} , \quad \hat{\lambda}_2^2 = \frac{n\sqrt{a_{11}/a_{22}}}{\sqrt{a_{11} a_{22}} + a_{12}} .$$

But $A = P^{-1} \times S$, so that

$$a_{11} = \frac{s_{11}}{1-\rho^2} ,$$

$$a_{22} = \frac{s_{22}}{1-\rho^2} ,$$

$$a_{12} = \frac{-s_{12}\rho}{1-\rho^2} ,$$

and hence

$$(2.5) \qquad \hat{\lambda}_1^2 = \frac{n(1-\rho^2)}{s_{11}(1-\rho r)} , \quad \hat{\lambda}_2^2 = \frac{n(1-\rho^2)}{s_{22}(1-\rho r)} ,$$

where $r = s_{12} / \sqrt{s_{11}s_{22}}$. Thus for the case $p = 2$,

$$(2.6) \qquad \underset{\lambda}{\text{Max}}\ p(S) = c(2,n;S) \frac{n^n e^{-n}}{(s_{11}s_{22})^{n/2}} \frac{(1-\rho^2)^{n/2}}{(1-\rho r)^n} .$$

For the usual unrestricted one-sample problem, the maximum of (2.6) with respect to ρ is achieved at $\hat{\rho} = r$, so that $\hat{\lambda}_1^2 = n/s_{11}$, $\hat{\lambda}_2^2 = n/s_{22}$, as expected. However, for other models we will have to obtain the maximum in (2.6) for ρ in a restricted domain.

3. Maximum likelihood estimators: Two bivariate populations.

In this section we obtain MLE of the parameters for the various models described in Section 1. We observe two independent sample covariance matrices, $\mathcal{L}(V) = W(\Sigma_1; 2,n)$ and $\mathcal{L}(W) = W(\Sigma_2; 2,m)$. The parameters in Σ_1 and Σ_2, namely, $(\sigma_{11},\sigma_{22},\rho_1)$ and $(\tau_{11},\tau_{22},\rho_2)$, respectively, depend on the structure assumed.

3.1 Model A. Here Σ_1 and Σ_2 have no parameters in common, so that each population covariance matrix is estimated by its respective sample covariance matrix, i.e.,

$$(3.1) \qquad \hat{\sigma}_{11} = \frac{v_{11}}{n} , \quad \hat{\sigma}_{22} = \frac{v_{22}}{n} , \quad \hat{\rho}_1 = r_1 ,$$

$$\hat{\tau}_{11} = \frac{w_{11}}{m} , \quad \hat{\tau}_{22} = \frac{w_{22}}{m} , \quad \hat{\rho}_2 = r_2 ,$$

and

$$(3.2) \qquad \underset{\Sigma_1,\Sigma_2}{\mathrm{Max}} \quad p(V,W) = \frac{c(2,n;V)c(2,m;W)\ n^n\ m^m\ e^{-(n+m)}}{(v_{11}v_{22})^{n/2}\ (w_{11}w_{22})^{m/2}\ (1-r_1^2)^{n/2}(1-r_2^2)^{m/2}} .$$

3.2 Model B. Here the variances in the two populations differ, but the correlation coefficient is common, i.e.,

$$\Sigma_1 = D_\sigma\ P\ D_\sigma \equiv D_\lambda^{-1}\ P\ D_\lambda^{-1} ,$$

$$\Sigma_2 = D_\tau\ P\ D_\tau \equiv D_\eta^{-1}\ P\ D_\eta^{-1} .$$

The joint density of V and W is

$$(3.3) \quad p(V,W) = c(2,n;V)c(2,m;W)(\lambda_1\lambda_2)^n(\eta_1\eta_2)^m|P|^{-(n+m)/2}\exp - \tfrac{1}{2}(\lambda A\lambda' + \eta B\eta'),$$

where $A = P^{-1} \times V$, $B = P^{-1} \times W$. Using the development of Section 2, we obtain

$$(3.4) \qquad \hat{\lambda}_1^2 = \frac{n(1-\rho^2)}{v_{11}(1-\rho r_1)} , \quad \hat{\lambda}_2^2 = \frac{n(1-\rho^2)}{v_{22}(1-\rho r_1)} ,$$

$$(3.5) \qquad \hat{\eta}_1^2 = \frac{m(1-\rho^2)}{w_{11}(1-\rho r_2)} , \quad \hat{\eta}_2^2 = \frac{m(1-\rho^2)}{w_{22}(1-\rho r_2)} .$$

This leads to the partial maximum

$$(3.6)\quad \underset{\rho,\lambda,\eta}{\mathrm{Max}}\ p(V,W) = \frac{c(2,n;V)c(2,m;W)e^{-(n+m)}n^n m^m}{(v_{11}v_{22})^{n/2}\,(w_{11}w_{22})^{m/2}}\ \underset{\rho}{\mathrm{Max}}\ \frac{(1-\rho^2)^{(n+m)/2}}{(1-\rho r_1)^n(1-\rho r_2)^m}\ .$$

We next show that the maximizer is

$$(3.7)\qquad \hat{\rho} = \frac{(n+m)(1+r_1r_2) - \sqrt{(n+m)^2(1-r_1r_2)^2 - 4nm(r_1-r_2)^2}}{2(mr_1 + nr_2)}\ .$$

Differentiation of (3.6) with respect to ρ leads to a quadratic, with roots $\hat{\rho} \equiv \rho_-$ given by (3.7) and $\hat{\rho} = \rho_+$ given by

$$\hat{\rho}_+ = \frac{(n+m)(1+r_1r_2) + \sqrt{(n+m)^2(1-r_1r_2)^2 - 4nm(r_1-r_2)^2}}{2(mr_1 + nr_2)}\ ;$$

For simplicity, write $\alpha = m/(n+m)$, $\beta = n/(n+m)$, so that $\alpha + \beta = 1$. The condition $\rho_+^2 \leq 1$ becomes, after simplification,

$$(3.8)\qquad (1+r_1r_2)\sqrt{(1-r_1r_2)^2 - 4\alpha\beta(r_1-r_2)^2} \leq 2(\alpha r_1^2 + \beta r_2^2) - (1+r_1^2r_2^2)\ .$$

The RHS of (3.8) may be positive or negative. If it is negative, then (3.8) cannot hold, so that ρ_- is the root in the interval $[-1,1]$. Suppose the RHS of (3.8) is positive then after squaring and collecting terms, we obtain

$$(3.9)\qquad \begin{aligned} 0 &\leq r_1^2r_2^2 + \alpha\,\beta(1+r_1r_2)^2(r_1-r_2)^2 + (\alpha r_1^2 + \beta r_2^2)^2 \\ &\quad - (1+r_1^2r_2^2)(\alpha r_1^2 + \beta r_2^2) \equiv g(\alpha)\ . \end{aligned}$$

We wish to assert that $g(\alpha) < 0$ for all α, so that (3.9) is violated, which completes the argument that ρ_+ is not in the interval $[-1,1]$.

To see this, note that $g(0) = -r_2^2(1-r_1^2)(1-r_2^2)$, $g(1) = -r_1^2(1-r_1^2)(1-r_2^2)$, and that $g'' = -2(r_1-r_2)^2(1-r_1^2)(1-r_2^2)$. Thus g is concave with $g(0) < 0$, $g(1) < 0$. Also

$$g' = -2(1-r_1^2)(1-r_2^2)(r_1-r_2)[r_2 + \alpha(r_1-r_2)] \ .$$

Since $g(\alpha) = 0$ at each of the extreme points $r_1 = r_2$, $r_1 = 1$, or $r_2 = 1$, we can omit these cases. Thus g has a maximum at $\hat{\alpha} = -r_2(r_1-r_2)$. But $g(\hat{\alpha}) = 0$, and hence $g(\alpha) < 0$ for $0 \le \alpha \le 1$.

Thus the MLE of $\lambda_i^2 = 1/\sigma_i^2$ and $\eta_i^2 = 1/\tau_i^2$ are given by (3.4) and (3.5) with $\hat{\rho}$ as in (3.7). From this we obtain

$$\max_{\rho,\lambda,\eta} p(V,W) = \frac{c(2,n;V)c(2,m;W)e^{-(n+m)}n^n m^m(1-\hat{\rho}^2)^{(n+m)/2}}{(v_{11}v_{22})^{n/2}(w_{11}w_{22})^{m/2}\ (1-\hat{\rho}r_1)^n(1-\hat{\rho}r_2)^m} \ . \tag{3.10}$$

If we use the z-transformation, i.e., $z_j = \frac{1}{2}\log(1+r_j)/(1-r_j)$, $j=1,2$, and form a weighted average,

$$\bar{z} = \frac{(n-3)z_1 + (m-3)z_2}{n+m-6} \equiv \mu z_1 + (1-\mu)z_2 \ , \tag{3.11}$$

then the inverse transformation yields

$$\bar{r} = \frac{e^{2\bar{z}}-1}{e^{2\bar{z}}+1} = \frac{(1+r_1)^{\mu}(1+r_2)^{1-\mu} - (1-r_1)^{\mu}(1-r_2)^{1-\mu}}{(1+r_1)^{\mu}(1+r_2)^{1-\mu} + (1-r_1)^{\mu}(1-r_2)^{1-\mu}} \ . \tag{3.12}$$

Although $\bar{r}$ is a reasonable estimator of ρ, it is not in general identical with $\hat{\rho}$. However, when $n=m$, i.e., $\mu = 1/2$, then

$$(3.13) \qquad \bar{r} = \hat{\rho} = \frac{1 + r_1 r_2 - \sqrt{(1-r_1^2)(1-r_2^2)}}{r_1 + r_2} \ .$$

It is easily seen that both $\hat{\rho}$ and $\bar{r}$ are consistent estimators of ρ. From

$$\sum_{j=1}^{2} \left(\frac{\partial\hat{\rho}}{\partial r_j}\bigg|_{\rho}\right)^2 E(r_j-\rho)^2 = \left(\frac{n}{n+m}\right)^2 \frac{(1-\rho^2)^2}{n} + \left(\frac{m}{n+m}\right)^2 \frac{(1-\rho^2)^2}{m}$$

$$= \frac{(1-\rho^2)^2}{n+m} \ ,$$

it follows that

$$(3.14) \qquad \mathcal{L}\left(\frac{\sqrt{n+m}\ (\hat{\rho}-\rho)}{1-\rho^2}\right) \xrightarrow[n,m]{} N(0,1) \ .$$

In a similar manner,

$$(3.15) \qquad \mathcal{L}\left(\frac{\sqrt{n+m-6}\ (\bar{r}-\rho)}{1-\rho^2}\right) \xrightarrow[n,m]{} N(0,1) \ .$$

3.3 Model C. Here the variances in the two populations are the same, but the correlations differ, i.e.,

$$\Sigma_j = D_\sigma P_j D_\sigma \equiv D_\lambda^{-1} P_j D_\lambda^{-1} \ , \quad j=1,2 \ .$$

The joint density function of V and W is

$$(3.16) \quad p(V,W) = c(2,n;V)c(2,m;W)(\lambda_1\lambda_2)^{n+m}|P_1|^{-n/2}|P_2|^{-m/2}\exp - \tfrac{1}{2}\lambda A\lambda' \ ,$$

where $A = (P_1^{-1} \times V) + (P_2^{-1} \times W)$. Using the development of Section 2, we obtain

$$(3.17)\qquad \hat{\lambda}_1^2 = \frac{(n+m)\sqrt{a_{22}/a_{11}}}{\sqrt{a_{11}a_{22}}+a_{12}}\ , \qquad \hat{\lambda}_2^2 = \frac{(n+m)\sqrt{a_{11}/a_{22}}}{\sqrt{a_{11}a_{22}}+a_{12}}\ ,$$

so that

$$(3.18)\qquad \max_{\lambda} p(V,W) = \frac{c(2,n;V)c(2,m;W)(n+m)^{n+m}e^{-(n+m)}}{(\sqrt{a_{11}a_{22}}+a_{12})^{n+m}(1-\rho_1^2)^{n/2}(1-\rho_2^2)^{m/2}}\ .$$

The problem remaining is to find

$$(3.19)\qquad \min_{\rho_1,\rho_2}\ [(n+m)\log(\sqrt{a_{11}a_{22}}+a_{12}) + \frac{n}{2}\log(1-\rho_1^2) + \frac{m}{2}\log(1-\rho_2^2)]\ ,$$

where

$$a_{jj} = \frac{v_{jj}}{1-\rho_1^2} + \frac{w_{jj}}{1-\rho_2^2}\ ,\quad j=1,2, \qquad a_{12} = -\frac{\rho_1 v_{12}}{1-\rho_1^2} - \frac{\rho_2 w_{12}}{1-\rho_2^2}\ .$$

The minimization leads to a cubic equation which can then be solved for any specific case. However, the general solution appears to be untractable.

3.4 Model D. This is the case when both populations have the same parameters. But if $\mathcal{L}(V) = W(\Sigma;\ 2,n)$ and $\mathcal{L}(W) = W(\Sigma;\ 2,m)$, with V and W independently distributed, then $\mathcal{L}(V+W) = W(\Sigma;\ 2,n+m)$. Consequently, $\hat{\Sigma} = (V+W)/(n+m)$, and

$$(3.20)\qquad \max_{\Sigma} p(V,W) = \frac{c(2,n;V)c(2,m;W)e^{-(n+m)}(n+m)^{n+m}}{|V+W|^{(n+m)/2}}\ .$$

4. Case of more than two populations.

We now observe k independent sample covariance matrices $\mathcal{L}(V_j) = W(\Sigma_j;\, 2, n_j)$, $j=1,2,\ldots,n$. In Model A, the generalization is immediate, namely, $\hat{\Sigma}_j = V_j/n_j$, and

$$(4.1) \qquad \underset{\Sigma_1,\ldots,\Sigma_k}{\text{Max}} \prod_1^k p(V_j) = \prod_1^k \frac{c(2,n_j;V_j)\, n_j^{n_j}\, e^{-n_j}}{[v_{11}^{(j)} v_{22}^{(j)} (1-r_j^2)]^{n_j/2}}$$

In Model B, the first step maximization leads to

$$(4.2) \qquad \underset{\rho, v_{11}^{(j)}, v_{22}^{(j)}}{\text{Max}} \prod_1^k p(V_j) = \prod_1^k \frac{c(2,n_j;V_j)\, n_j^{n_j} e^{-n_j}}{(v_{11}^{(j)}\, v_{22}^{(j)})^{n_j/2}} \; \underset{\rho}{\text{Max}} \frac{(1-\rho^2)^{\Sigma n_j/2}}{\prod_1^k (1-\rho r_j)^{n_j}} .$$

Setting the derivative of the logarithm equal to zero yields the equation

$$\frac{n\rho}{1-\rho^2} = \frac{n_1 r_1}{1-\rho r_1} + \cdots + \frac{n_k r_k}{1-\rho r_k} ,$$

which is of the k-th degree. Although the solution cannot be obtained in closed form, it can be carried out numerically.

As in the case of the two populations, Model C appears to be untractable. Model D can be carried out very simply and the generalization is immediate.

5. Tests of hypotheses.

There are four tests which can be considered: (i) test for the equality of two correlation coefficients when the variances in the two populations are known to be unrestricted or the same, (ii) test for the equality of the two pairs of variances when the correlation coefficients are unrestricted or the same. The two cases for the first hypothesis are Model B versus Model A, and Model D versus Model C. The two cases for the second hypothesis are Model C versus Model A, and Model D versus Model B.

Thus, to test $H: \rho_1 = \rho_2$ when $\sigma_{11}, \sigma_{22}, \tau_{11}, \tau_{22}$ are unrestricted, we obtain the likelihood ratio test as the ratio of (3.2) to (3.10), namely

$$L = \frac{(1-\hat{\rho}^2)^{(n+m)/2}(1-r_1^2)^{n/2}(1-r_2^2)^{m/2}}{(1-\hat{\rho}r_1)^n\,(1-\hat{\rho}r_2)^m}\,, \tag{5.1}$$

where $\hat{\rho}$ is given by (3.7). To carry out the test we use the large sample result that $-2 \log L$ is distributed approximately as a chi-square variate with one degree of freedom.

To test that the pairs of variances are equal when the correlation coefficients are known to be equal, we use the ratio of (3.20) to (3.10), namely,

$$L = \frac{(n+m)^{n+m}\,(v_{11}v_{22})^{n/2}(w_{11}w_{22})^{m/2}\,(1-\hat{\rho}r_1)^n(1-\hat{\rho}r_2)^m}{n^n\,m^m\,|V+W|^{(n+m)/2}\,(1-\hat{\rho}^2)^{(n+m)/2}}\,,$$

where $\hat{\rho}$ is given by (3.7). As before, $\mathcal{L}(-2 \log L) \to \chi_1^2$.

6. Intraclass correlation models.

Suppose $\mathcal{L}(V) = W(\Sigma; p,n)$, where $\sigma_{ii} = \sigma^2$, $\sigma_{ij} = \sigma^2\rho (i \neq j)$, i.e., $\Sigma = \sigma^2[(1-\rho)I + \rho e'e]$, where $e = (1,\ldots,1)$. By an orthogonal transformation $V \to \Gamma V \Gamma' \equiv W$, with the first row of Γ equal to $e/\sqrt{p}$, $\Sigma \to \text{diag}\,(\alpha,\beta,\ldots,\beta)$, where $\alpha = \sigma^2[1+(p-1)\rho]$, $\beta = \sigma^2(1-\rho)$. Consequently, $(w_{11}, \sum_2^p w_{ii}) \equiv (u,v)$ is a sufficient statistic for (α,β). It is straightforward to show that u and v are independently distributed, $\mathcal{L}(u/\alpha) = \mathcal{L}(\chi^2_\mu)$, $\mathcal{L}(v/\beta) = \mathcal{L}(\chi^2_\nu)$, where μ and ν are functions of n and m, respectively, and we consider this as our starting point. (Various models concerning the means of the underlying normal population are discussed by Olkin and Pratt [4]. In most cases ν and μ are of the order pn and n, respectively.)

Since $\hat{\alpha} = u/\mu$, $\hat{\beta} = v/\nu$, we have that

$$(6.1) \qquad \hat{\rho} = \frac{\hat{\alpha}-\hat{\beta}}{\hat{\alpha} + (p-1)\hat{\beta}}, \qquad \hat{\sigma}^2 = \frac{\hat{\alpha} + (p-1)\hat{\beta}}{p}.$$

6.1 Model a. The population parameters are σ^2, ρ_1 and τ^2, ρ_2, so that $\alpha_1,\beta_1,\alpha_2,\beta_2$ are unrestricted. Thus if $\mathcal{L}(u_j/\alpha_j) = \mathcal{L}(\chi^2_{\mu_j})$, $\mathcal{L}(v_j/\beta_j) = \mathcal{L}(\chi^2_{\nu_j})$, $j=1,2$, we have ρ_1, σ^2, ρ_2, τ^2 given by (6.1) with the appropriate subscripts.

6.2 Model b. In this case, the variances differ, but the correlations are equal, i.e., the parameters are σ^2, τ^2, ρ. It is easily seen that

$$\hat{\sigma}^2 = \left(\frac{u_1}{1+(p-1)\rho} + \frac{v_1}{1-\rho}\right) \Big/ (\mu_1 + \nu_1) ,$$

$$\hat{\tau}^2 = \left(\frac{u_2}{1+(p-1)\rho} + \frac{v_2}{1-\rho}\right) \Big/ (\mu_2 + \nu_2) ,$$

and hence

$$(6.2) \quad \underset{\sigma^2,\tau^2}{\text{Max}}\; p(u_1,u_2,v_1,v_2) = c \prod_{j=1}^{2} \frac{(1+(p-1)\rho]^{\nu_j/2}\,[1-\rho]^{\mu_j/2}}{[u_j + v_j + \rho(v_j(p-1)-u_j)]^{(\mu_j+\nu_j)/2}} ,$$

where

$$c = \prod_{j=1}^{2} c(1,\mu_j;u_j)\, c(1,\nu_j;v_j) e^{-\frac{1}{2}(\mu_j+\nu_j)} (\mu_j+\nu_j)^{(\mu_j+\nu_j)/2} .$$

If we define $w_j = v_j/(u_j+v_j)$, and $x = p\rho/(1-\rho)$, the problem reduces to

$$(6.3) \quad \underset{\rho}{\text{Max}} \prod_{j=1}^{2} \frac{[1+(p-1)\rho]^{\nu_j/2}[1-\rho]^{\mu_j/2}}{[1+(pw_j-1)\rho]^{(\mu_j+\nu_j)/2}}$$

$$= \underset{x}{\text{Max}} \prod_{j=1}^{2} \frac{(1+x)^{\nu_j/2}}{(1+w_j x)^{(\mu_j+\nu_j)/2}} \equiv \underset{x}{\text{Max}}\; \psi(x) .$$

Setting the derivatives of the logarithm equal to zero yields

$$\frac{(\nu_1+\nu_2)}{1+x} - \frac{(\mu_1+\nu_1)w_1}{1+w_1 x} - \frac{(\mu_2+\nu_2)w_2}{1+w_2 x} = 0 ,$$

or

$$(6.4) \quad x^2 - gx - h = 0 ,$$

where

$$(6.5)\qquad g = \frac{\nu_2-\mu_1}{\mu_1+\mu_2}\,\frac{1}{w_2} + \frac{\nu_1-\mu_2}{\mu_1+\mu_2}\,\frac{1}{w_1} - \left(1+\frac{\nu_1+\nu_2}{\mu_1+\mu_2}\right),$$

$$h = \frac{\nu_1+\nu_2}{\mu_1+\mu_2} - \frac{\nu_1+\mu_1}{\mu_1+\mu_2}\,\frac{1}{w_2} - \frac{\nu_2+\mu_2}{\mu_1+\mu_2}\,\frac{1}{w_1}.$$

Since it is easy to see that there exist the solution of(6.3), at least one of

$$(6.6)\qquad x_+ = \frac{1}{2}\,[g+\sqrt{g^2+4h}]\,,\quad x_- = \frac{1}{2}\,[g-\sqrt{g^2+4h}]$$

is our maximizer.

We now show that x_- cannot be the maximizer. First note that $\psi(x)$ is defined over $(-1, \infty)$, and $\psi'(0) = \frac{1}{2}h(\mu_1+\mu_2)w_1w_2$. It is easy to show that if $h > 0$, $\psi'(x) > 0$ for all x in $(-1,0)$ and if $h < 0$, $\psi'(x) < 0$ for all x in $[0, \infty]$, and if $h < 0$, it must be in $(-1,0)$. However, wehave the following signs of x_+ and x_- :

		x_+	x_-
$h > 0$	$g > 0$	> 0	< 0
	$g < 0$	> 0	< 0
$h < 0$	$g > 0$	> 0	> 0
	$g < 0$	< 0	< 0

From this, it is seen that in the case $h > 0$, x_- is not the solution. The case $h < 0$, $g > 0$, is impossible, because the maximizer is in $[-1,0]$, where both x_+ and x_- are positive. It remains to prove that

$x_- \le -1$, if $h < 0$. To do this let us assume that x_- is the solution, i.e., $-1 < x_- \le 0$. Then $\sqrt{g^2+4h} < g+2$. After squaring and simplifying, we obtain $0 < g-h+1$. But this is impossible, because

$$g+h-1 = -\frac{(\nu_1+\nu_2)(1-w_1)(1-w_2)}{(\mu_1+\mu_2)w_1w_2} < 0 .$$

Thus, the maximizer of (6.3) is $x = x_+$, and hence we have, as the maximum likelihood estimator of ρ,

$$\hat{\rho} = \frac{\hat{x}}{\hat{x}+p} = \frac{gp - 2h + p\sqrt{g^2+4h}}{2(p^2 + gp - h)} , \tag{6.7}$$

where p, g, and h are defined by (6.5).

In reliability models, functions such as $\gamma \equiv [1+k\rho]/[1+(p-1)\rho]$ arise. In the case of Model b, we obtain

$$\hat{\gamma} = \frac{p+(k+1)\hat{x}}{p(1+\hat{x})} ,$$

where $\hat{x}$ is defined by (6.3).

6.3 Asymptotic Distributions. We first show that

$$\mathcal{L}(\hat{x}) \to N(\phi-1, \frac{2\phi^2\kappa_1\kappa_2}{\kappa_1+\kappa_2}) , \tag{6.8}$$

where $\phi = [1+(p-1)\rho]/(1-\rho)$, $\kappa_i = \mu_i^{-1} + \nu_i^{-1}$, $i=1,2$, and the limiting process is taken as $\mu_i \to \infty$, $\nu_i \to \infty$, with μ_i/ν_i constant. From (6.8) it will follow that

$$\mathcal{L}(\hat{\rho}) \to N(\rho, \frac{2\kappa_1\kappa_2}{\kappa_1+\kappa_2} \frac{(1-\rho)^2[1+(p-1)\rho]^2}{p^2}) \tag{6.9}$$

Asymptotic distributions for other functions of ρ can be easily obtained from (6.8) or (6.9). For example, if $\hat{\phi} = [1+(p-1)\hat{\rho}]/(1-\hat{\rho})$, then $\mathcal{L}(\hat{\phi}) \rightarrow \mathcal{N}(\phi, 2\phi^2\kappa_1\kappa_2/(\kappa_1+\kappa_2))$.

To carry out the computations, recall from (6.5) and (6.6) that

$$2\hat{x} = g + \sqrt{g^2 + 4h} \ ,$$

where

$$g \equiv g(y_1,y_2) = \frac{\nu_2-\mu_1}{\mu_1+\mu_2} y_2 + \frac{\nu_1-\mu_2}{\mu_1+\mu_2} y_1 - (1+\frac{\nu_1+\nu_2}{\mu_1+\mu_2}) \ ,$$

$$h \equiv h(y_1,y_2) = \frac{\nu_1+\nu_2}{\mu_1+\mu_2} - \frac{\nu_1+\mu_1}{\mu_1+\mu_2} y_2 - \frac{\nu_2+\mu_2}{\mu_1+\mu_2} y_1 \ ,$$

where $y_1 = 1/w_1$, $y_2 = 1/w_2$.

Since u_i and v_j are independent, we have

$$\mathcal{L}(u_i) \rightarrow \mathcal{N}(\alpha_i\mu_i, 2\alpha_i^2\mu_i) \ , \quad \mathcal{L}(v_i) \rightarrow \mathcal{N}(\beta_i\nu_i, 2\beta_i^2\nu_i)$$

as $\mu_i \rightarrow \infty$ and $\nu_i \rightarrow \infty$. Hence, for $y_i = 1+u_iv_i^{-1}$, it follows that

$$\mathcal{L}(y_i) \rightarrow N(1+\frac{\mu_i}{\nu_i}\phi \ , \ 2(\frac{\mu_i}{\nu_i})^2 \phi^2\kappa_i), \ i=1,2 \ ,$$

where $\phi = \alpha_i/\beta_i = [1+(p-1)\rho]/(1-\rho)$. We also note that y_1 and y_2 are independent. By expanding $g(y_1,y_2)$ and $h(y_1,y_2)$ about the point (y_{10},y_{20}), $y_{i0} = 1 + \frac{\mu_i}{\nu_i}\phi$, we find that $2\hat{x}$ is asymptotically normally distributed with mean

$$2x_o \equiv g_o + \sqrt{g_o^2 + 4h_o} = [-2+(1-m)\phi] + \sqrt{(1+m)^2\phi^2}$$

$$= 2(\phi-1) \ ,$$

where $m = (\nu_1^{-1}+\nu_2^{-1})/(\mu_1^{-1}+\mu_2^{-1})$.

The asymptotic variance is

$$v_o = \left(\frac{\partial(2\hat{x})}{\partial y_1}\right)^2 \kappa_1 \, V(y_1) + \left(\frac{\partial(2\hat{x})}{\partial y_2}\right)^2 \kappa_2 \, V(y_2) ,$$

evaluated at (y_{10}, y_{20}). But

$$\left.\frac{\partial(2\hat{x})}{\partial y_1}\right|_{y_o} = \left(1+ \frac{g_o}{\sqrt{g_o^2 + 4h_o}}\right) \frac{\partial g}{\partial y_1} + \frac{2}{\sqrt{g_o^2 + 4h_o}} \frac{\partial h}{\partial y_1}$$

$$= \frac{2(\varphi-1)}{(1+m)\varphi} \frac{\nu_1+\nu_2}{\mu_1+\mu_2} + \frac{2}{(1+m)\varphi} \frac{1}{\mu_1+\mu_2} [(\nu_1-\mu_2)+(\nu, +\nu_2)(\mu_2\nu_2^{-1})\varphi]$$

$$= \frac{2}{1+m} \frac{\nu_2+\mu_2}{\mu_1+\mu_2} \frac{\nu_1}{\nu_2} ,$$

$$\left.\frac{\partial(2\hat{x})}{\partial y_2}\right|_{y_o} = \frac{2}{1+m} \frac{\nu_1+\mu_1}{\mu_1+\mu_2} \frac{\nu_2}{\nu_1} ,$$

and hence we obtain after simplification,

$$v_o = \frac{8\varphi^2 \kappa_1 \kappa_2}{\kappa_1+\kappa_2} .$$

REFERENCES

[1] Bellman, Richard, (1960). Introduction to Matrix Analysis, McGraw-Hill, New York.

[2] Durost, Walter, N., (1932). Children's Collecting Activity Related to Social Factors. Bureau of Publications, Teachers College, Columbia University, New York.

[3] Marshall, Albert, W. and Olkin, Ingram, (1960). "Multivariate Chebyshev inequalities," Ann. Math. Statist. 31, 1001-1014.

[4] Olkin, Ingram and Pratt, John W., (1958). "Unbiased estimation of certain correlation coefficients," Ann. Math. Statist. 29, 201-211.

[5] Walker, Helen M., and Lev, Joseph, (1953). Statistical Inference, Henry Holt and Company, New York.

ASYMPTOTIC DISTRIBUTION OF FUNCTIONS OF A CORRELATION MATRIX

by

Ingram Olkin and Minoru Siotani

1. Introduction.

Although the joint distribution of the product-moment correlation coefficients in the central case, i.e., when the variates are independent, is an old result, very few functions of the correlations have been studied. Most well known are the partial and multiple correlation coefficients, for which the non-central distributions have been studied extensively. However, the non-central joint distribution of the product moment correlations or the joint distribution of product moment, partial, and/or multiple correlations has been almost completely neglected. This is due primarily to the fact that the exact distributions are clearly untractable. On the other hand, because the correlations are functions of moments, the usual delta method yields asymptotic normality, and the determination of the asymptotic covariance matrix is frequently quite manageable.

The purpose of this paper is to obtain a number of asymptotic distributions for the non-central case, in the hope that this will shed light on the behavior of these statistics when they serve as estimators or for tests of hypothesis.

Some preliminaries are discussed in Section 2; in particular, the basic Theorem used for obtaining asymptotic joint distributions is presented. In the remaining sections the asymptotic joint distribution of the following statistics in the non-central case are presented:

(i) Elements of the correlation matrix.

(ii) Product moment and partial correlation coefficients.

(iii) Product moment and multiple correlation coefficients.

(iv) Determinant of the correlation matrix.

(v) Determinants of step-down correlation matrices.

(vi) Determinants of certain submatrices.

(vii) Likelihood ratio criterion for testing independence of sets of variables.

2. Preliminaries.

In order to derive many of the asymptotic distributions, we make repeated use of the following Theorem, which is a straightforward generalization of a result of Anderson [1, Theorem 4.2.5].

Theorem 2.1. Let $u(n) = (u_1(n),\dots,u_m(n))$ be a random vector such that $\operatorname{plim}_{n\to\infty} u(n) = b \equiv (b_1,\dots,b_m)$ and $\sqrt{n}\,(u(n)-b)$ is asymptotically normally distributed with mean vector 0 and covariance matrix ψ. If $y(n) \equiv (y_1(n),\dots,y_k(n)) = (f_1,\dots,f_k)$, $k \le m$, where $f_i \equiv f_i(u(n))$ are functions of $u(n)$ having first and second derivatives in a neighborhood of $u(n) = b$, then

$$\mathcal{L}(\sqrt{n}\,\{y(n)-f(b)\}) \to \mathcal{N}(0, A\,\psi\,A')$$

where A is a $k \times m$ matrix with elements $a_{ij} = \left.\frac{\partial f_i(u)}{\partial u_j(n)}\right|_{u=b}$.

Let X be a $k \times N$ matrix with columns which are independently and identically distributed, each having a $N(\mu,\Sigma)$ distribution, from which we obtain that $(\bar{x},S)$ is a sufficient statistic for (μ,Σ), where $\bar{x} = (\bar{x}_1,\ldots,\bar{x}_k)$ is the mean vector, $S = (XX'-N\bar{x}'\bar{x})/n$ is the sample covariance matrix, $n = N-1$. It is well-known that $\operatorname*{plim}_{n\to\infty} S = \Sigma$ and the limiting distribution of $\sqrt{n}\ \{S-\Sigma\}$, i.e., the joint distribution of the s_{ij}, is normal with means 0 and covariances

$$\text{(2.1)} \qquad \operatorname{Cov}(s_{ij},s_{k\ell}) = \frac{1}{n}[\sigma_{ik}\,\sigma_{j\ell} + \sigma_{i\ell}\,\sigma_{jk}] \ .$$

If we let $U = (u_{ij})$, $u_{ij} = s_{ij}/\sqrt{\sigma_{ii}\sigma_{jj}}$, then the limiting distribution of $\sqrt{n}\ \{U-P\}$ is normal with means 0 and covariances

$$\text{(2.2)} \qquad \operatorname{Cov}(u_{ij},u_{k\ell}) = \frac{1}{n}[\rho_{ik}\,\rho_{j\ell} + \rho_{i\ell}\,\rho_{jk}] \ ,$$

where $P = (\rho_{ij})$ is the matrix of correlation coefficients. By arranging the independent elements of U and P in a row as

$$u = (u_{11},\ldots,u_{kk},u_{12},u_{13},\ldots,u_{1k},u_{23},\ldots,u_{2k},\ldots,u_{k-1,k}) \ ,$$
$$\rho = (1,\ldots,\ 1\ ,\ \rho_{12},\rho_{13},\ldots,\rho_{1k},\rho_{23},\ldots,\rho_{2k},\ldots,\rho_{k-1,k}) \ ,$$

we have that $\sqrt{n}\ \{u-\rho\}$ is asymptotically normally distributed with mean vector 0 and covariances given by (2.2).

Since the statistics of concern are functions of the elements of u, we can consider the asymptotic distribution of u as our starting point. To obtain a specific distribution, we invoke Theorem 2.1, so that we need only obtain the asymptotic covariance structure.

To distinguish between exact and limiting covariances, we write Cov as L-Cov, respectively.

3. Asymptotic joint distribution of correlation coefficients.

By defining $r_{ij} = f_{ij}(u) = u_{ij}/\sqrt{u_{ii}u_{jj}}$,

$$r = (r_{12}, r_{13}, \ldots, r_{1n}, r_{23}, \ldots, r_{2k}, \ldots, r_{k-1,k}) ,$$

$$\rho = (\rho_{12}, \rho_{13}, \ldots, \rho_{1k}, \rho_{23}, \ldots, \rho_{2k}, \ldots, \rho_{k-1,k}) ,$$

we immediately obtain that $\sqrt{n}\,\{r-\rho\}$ is asymptotically normally distributed with means zero and covariance matrix $A\,\psi\,A'$, where A is the matrix of partial derivatives, i.e., $(\partial f_{ij}/\partial u_{k\ell})$ and ψ is proportional to the matrix of covariances $\mathrm{Cov}(u_{ij}, u_{k\ell})$. The result of this calculation is (with 1, 2, 3, 4 used for i,j,k,ℓ)

$$\text{(3.1)} \qquad \text{L-Cov}(r_{12}, r_{34}) = \frac{1}{n}\Big[\frac{1}{2}\,\rho_{12}\rho_{34}(\rho_{13}^2 + \rho_{14}^2 + \rho_{23}^2 + \rho_{24}^2)$$

$$+ \rho_{13}\rho_{24} + \rho_{14}\rho_{23} - (\rho_{12}\rho_{13}\rho_{14} + \rho_{21}\rho_{23}\rho_{24} + \rho_{31}\rho_{32}\rho_{34} + \rho_{41}\rho_{42}\rho_{43})\Big]$$

This simplifies when the subscripts are duplicated, e.g.,

$$\text{(32)} \qquad \text{L-Cov}(r_{12}, r_{13}) = \frac{1}{n}\Big[\frac{1}{2}(2\rho_{23} - \rho_{12}\rho_{13})(1 - \rho_{12}^2 - \rho_{13}^2 - \rho_{23}^2) + \rho_{23}^3\Big] ,$$

$$\text{L-Var}(r_{12}) = \frac{1}{n}(1 - \rho_{12}^2)^2 .$$

These formulas hold for any k, however, when $k = 3$, the result reduces since there must be a duplication of subscripts. Thus $\sqrt{n}\,\{(r_{12}, r_{13}, r_{23}) - (\rho_{12}, \rho_{13}, \rho_{23})\}$ is asymptotically normally distributed with means 0 and

$$\text{L-Var}(r_{12}) = \frac{1}{n}(1-\rho_{12}^2)^2 ,$$

$$\text{L-Cov}(r_{12},r_{13}) = \frac{1}{n}[\tfrac{1}{2}(2\rho_{32}-\rho_{12}\rho_{13})(1-\rho_{12}^2-\rho_{13}^2-\rho_{23}^2) + \rho_{23}^3] ,$$

with similar expressions for other subscripts.

In the special case when $\rho_{ij} = \rho$, the asymptotic covariances become

$$\text{L-Var}(r_{12}) = \frac{1}{n}(1-\rho^2)^2 ,$$

$$\text{L-Cov}(r_{12},r_{13}) = \frac{1}{n}[\tfrac{\rho}{2}(1-\rho)^2(2+3\rho)] ,$$

$$\text{L-Cov}(r_{12},r_{34}) = \frac{1}{n}[2\rho^2(1-\rho)^2] .$$

4. Some special results for the trivariate case.

In this section we are concerned with two special joint distributions, namely, that of $(r_{12},r_{12\cdot 3})$ and $(r_{12},r_{23},r_{1(23)})$, where $r_{12\cdot 3},r_{1(23)}$ are the partial and multiple correlation coefficients, respectively.

Define $f_1(r) = r_{12}$, $f_2(r) = (r_{12}-r_{13}r_{23})/\sqrt{(1-r_{13}^2)(1-r_{23}^2)}$, then using the asymptotic distribution of (r_{12},r_{13},r_{23}) with Theorem 2.1, we obtain

Theorem 4.1. The limiting joint distribution of $\sqrt{n}\ [(r_{12},r_{12\cdot 3})-(\rho_{12},\rho_{12\cdot 3})]$ is normal with mean vector $(0,0)$ and covariances

$$\text{L-Var}(r_{12}) = \frac{1}{n}(1-\rho_{12}^2)^2 , \quad \text{L-Var}(r_{12\cdot 3}) = \frac{1}{n}(1-\rho_{12\cdot 3}^2)^2 ,$$

$$\text{L-Cov}(r_{12},r_{12\cdot 3}) = \frac{1}{n}[\tfrac{1}{2}\,\rho_{12}\rho_{13}\rho_{23\cdot 1}\,\frac{(1-\rho_{12}^2-\rho_{13}^2)}{1-\rho_{13}^2} + \tfrac{1}{2}\,\rho_{12}\rho_{23}\rho_{13\cdot 2}\,\frac{(1-\rho_{12}^2-\rho_{23}^2)}{1-\rho_{23}^2}$$
$$+ (1-\rho_{12}^2)^2/\sqrt{(1-\rho_{13}^2)(1-\rho_{23}^2)}\ ] .$$

In the special case when $\rho_{13} = \rho_{23} = 0$, i.e., when x_3 is independent x_1 and x_2, $\text{L-Cov}(r_{12}, r_{12\cdot 3}) = \frac{1}{n}(1-\rho_{12}^2)^2$, so that the covariance matrix is singular, as expected.

To obtain the joint distribution of r_{12}, r_{23} and $r_{1(23)}$, where

$$r_{1(23)}^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1-r_{23}^2},$$

define $f_1(r) = r_{12}$, $f_2(r) = r_{23}$, $f_3(r) = r_{1(23)}$. This leads to

Theorem 4.2. The limiting joint distribution of

$$\sqrt{n}\,\{(r_{12}, r_{23}, r_{1(23)}) - (\rho_{12}, \rho_{23}, \rho_{1(23)})\}$$

is normal with mean vector $(0,0,0)$ and covariance matrix proportional to

$$\begin{pmatrix} \mathrm{Var}(r_{12}) & \mathrm{Cov}(r_{12},r_{13}) & \rho_{1(23)}^{-1}\left[\mathrm{Var}(r_{12})\frac{\rho^{12}}{\rho^{11}} + \mathrm{Cov}(r_{12},r_{13})\frac{\rho^{13}}{\rho^{11}} - \mathrm{Cov}(r_{12},r_{23})\frac{\rho^{12}\rho^{13}}{\rho^{11}\rho^{11}}\right] \\ - & \mathrm{Var}(r_{23}) & \rho_{1(23)}^{-1}\left[\mathrm{Cov}(r_{12},r_{13})\frac{\rho^{12}}{\rho^{11}} + \mathrm{Cov}(r_{13},r_{23})\frac{\rho^{13}}{\rho^{11}} - \mathrm{Var}(r_{23})\frac{\rho^{12}\rho^{13}}{\rho^{11}\rho^{11}}\right] \\ - & - & [1-\rho_{1(23)}^2]^2 \end{pmatrix}$$

where the blank entries below the diagonal are determined by the condition of symmetry, and where $P^{-1} = (\rho^{ij})$.

To see this, note that the asymptotic covariance matrix is $A \psi A'$, where

$$A = \begin{pmatrix} \frac{\partial f_1}{\partial r_{12}} & \frac{\partial f_1}{\partial r_{13}} & \frac{\partial f_1}{\partial r_{23}} \\ \frac{\partial f_2}{\partial r_{12}} & \frac{\partial f_2}{\partial r_{13}} & \frac{\partial f_2}{\partial r_{23}} \\ \frac{\partial f_3}{\partial r_{12}} & \frac{\partial f_3}{\partial r_{13}} & \frac{\partial f_3}{\partial r_{23}} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ \frac{\rho^{12}}{\rho_{1(23)}\rho^{11}} & \frac{\rho^{13}}{\rho_{1(23)}\rho^{11}} & \frac{-\rho^{12}\rho^{13}}{\rho_{1(23)}(\rho^{11})^2} \end{pmatrix}$$

and ψ is the asymptotic covariance matrix of $\sqrt{n}\,(r_{12}, r_{13}, r_{23})$ given by (3.2).

5. Asymptotic distribution of the determinant of the correlation matrix.

Let $R = (r_{ij})$ be a sample correlation matrix calculated from a sample of size $N \equiv n+1$ drawn from a k-variate normal distribution with population correlation matrix $P = (\rho_{ij})$. We now obtain the non-central asymptotic distribution of the determinant, $|R|$. This statistic also arises as the likelihood ratio criterion for testing for independence in a normal population, so that the following result permits a determination of asymptotic power.

Theorem 5.1. The limiting distribution of $\sqrt{n}\,(|R| - |P|)$ is normal with mean 0 and variance $2|P|^2\,(\text{tr.}\,P^2 - k)$.

Proof. By Theorem 2.1 the limiting distribution of $\sqrt{n}\,\{|R|-|P|\}$ is normal with mean 0, and we need only evaluate the variance, i.e.,

$$(5.1)\qquad E\{\sum_{i<j} \frac{\partial|P|}{\partial\rho_{ij}} (r_{ij}-\rho_{ij})\}^2 = \sum_{i,j,\ell,m} |P|^2\, \rho^{ij}\rho^{\ell m}\, \mathrm{Cov}(r_{ij},r_{\ell m}) ,$$

where $r_{ii} = 1$, so that $\mathrm{Cov}(r_{ii},r_{jk}) = 0$, and $P^{-1} = (\rho^{ij})$.

From (3.1) we have

$$\text{L-Cov}(r_{ij},r_{\ell m}) = \frac{1}{n}[\frac{1}{2}\, \rho_{ij}\rho_{\ell m}(\rho_{i\ell}^2 + \rho_{im}^2 + \rho_{j\ell}^2 + \rho_{jm}^2) + \rho_{i\ell}\rho_{jm} + \rho_{im}\rho_{j\ell}$$

$$-(\rho_{ij}\rho_{i\ell}\rho_{im} + \rho_{ji}\rho_{j\ell}\rho_{jm} + \rho_{\ell i}\rho_{\ell j}\rho_{\ell m} + \rho_{mi}\rho_{mj}\rho_{m\ell})] .$$

Because of invariance under permutations of the subscripts, we need only evaluate

$$\sum_{i,j,\ell,m} |P|^2\, \rho^{ij}\rho^{\ell m}\, \rho_{ij}\rho_{\ell m}\rho_{i\ell}^2 = \sum_{\alpha,\beta} \rho_{\alpha\beta}^2\; |P|^2 = |P|^2 \operatorname{tr} P^2 ,$$

$$\sum_{i,j,\ell,m} |P|^2\, \rho^{ij}\rho^{\ell m}\, \rho_{i\ell}\rho_{jm} = k|P|^2 ,$$

$$\sum_{i,j,\ell,m} |P|^2\, \rho^{ij}\rho^{\ell m}\rho_{ij}\rho_{i\ell}\rho_{im} = k|P|^2 ,$$

from which the result follows immediately.

Remark. In the null case when $P = I$, the variance is zero since $\operatorname{tr} P^2 = k$. Hence, as far as terms of order n^{-1} are concerned, $\sqrt{n}\,\{|R|-1\}$ is asymptotically normal with zero variance. It is easy to check that the variance of $|R|$ is of order n^{-2}; in fact, from the null distribution of R, (e.g., see Anderson [1], p. 175)

$$E|R|^h = \prod_{1}^{k} \left[\frac{\Gamma(\frac{1}{2} n)\ \Gamma(\frac{1}{2}(n-i+1)+h)}{\Gamma(\frac{1}{2} n+h)\ \Gamma(\frac{1}{2}(n-i+1))} \right] .$$

In particular,

$$E|R| = 1 - \frac{k(k-1)}{2n} + O(n^{-2}) ,$$

$$E|R|^2 = 1 - \frac{(n+1)k(k-1)}{n(n+2)} + O(n^{-2}) ,$$

and hence $\text{L-Var}|R| = O(n^{-2})$. A similar result holds for the asymptotic distribution of the multiple correlation coefficients.

6. Asymptotic joint distribution of determinants of step-down correlation matrices.

Denote the step-down sample and population correlation matrices by $R_{(m)} = (r_{ij})$ and $P_{(m)} = (p_{ij})$, $i,j=1,\ldots,m$, respectively. (Note that $R_{(1)} = 1$ and $R_{(k)} = R$.) As before, it is clear that

$$\sqrt{n}\ [(|R_{(2)}|,\ldots,|R_{(k)}|) - (|P_{(2)}|,\ldots,|P_{(k)}|)]$$

is asymptotically normal with mean vector $(0,\ldots,0)$. To evaluate the covariance matrix, let $P^{-1}_{(\alpha)} = (\rho^{ij}_{(\alpha)})$, $\alpha=1,\ldots,k$; $i,j \le \alpha$. Since

$$\frac{\partial|P_{(i)}|}{\partial\rho_{su}} = 2\ \rho^{su}_{(i)}\ |P_{(i)}| ,$$

the asymptotic variance of $|R_{(i)}|$ is

$$4 \sum_{s<u} \sum_{v<w} |P_{(i)}|^2 \rho^{su}_{(i)} \rho^{vw}_{(i)} \operatorname{Cov}(r_{su}, r_{vw})$$

which is the same as (5.1), so that

$$\text{L-Var}(|R_{(j)}|) = \frac{2}{n}|P_{(j)}|^2(\operatorname{tr} P^2_{(j)} - j) \ , \ j=2,\ldots,k \ .$$

The covariance between $|R_{(i)}|$ and $|R_{(j)}|$ can be similarly evaluated. In fact, with $i < j$,

$$(6.1) \quad \text{L-Cov}(|R_{(i)}|, |R_{(j)}|) = \sum_{s,u=1}^{i} \sum_{v,w=1}^{j} |P_{(i)}||P_{(j)}| \rho^{su}_{(i)} \rho^{vw}_{(j)} \operatorname{Cov}(r_{su}, r_{vw})$$

$$= \frac{1}{n} \sum_{s,u=1}^{i} \sum_{v,w=1}^{j} |P_{(i)}||P_{(j)}| \rho^{su}_{(i)} \rho^{vw}_{(j)} \{\frac{1}{2} \rho_{su}\rho_{vw}(\rho^2_{sv}+\rho^2_{sw}+\rho^2_{uv}+\rho^2_{uw})$$

$$+ \rho_{sv}\rho_{uw}+\rho_{sw}\rho_{uv}-(\rho_{su}\rho_{sv}\rho_{sw}+\rho_{us}\rho_{uv}\rho_{uw}+\rho_{vs}\rho_{vu}\rho_{vw}+\rho_{ws}\rho_{wu}\rho_{wv})\}$$

$$= \frac{2}{n} \sum_{s,u=1}^{i} \sum_{v,w=1}^{j} |P_{(i)}||P_{(j)}| \rho^{su}_{(i)} \rho^{vw}_{(j)} \{\rho_{su}\rho_{vw}\rho^2_{sv}+\rho_{sv}\rho_{uw}-\rho_{su}\rho_{sv}\rho_{sw}-\rho_{vs}\rho_{vu}\rho_{vw}\} .$$

Since

$$(6.2) \quad \sum_{s,u=1}^{i} \sum_{v,w=1}^{j} |P_{(i)}||P_{(j)}| \rho^{su}_{(i)} \rho^{vw}_{(j)} \rho_{su}\rho_{vw}\rho^2_{sv} = |P_{(i)}||P_{(j)}| \sum_{s=1}^{i} \sum_{v=1}^{j} \rho^2_{sv} ,$$

$$(6.3)\qquad \sum_{s,u=1}^{i}\ \sum_{v,w=1}^{j}|P_{(i)}||P_{(j)}|\rho^{su}_{(i)}\rho^{vw}_{(j)}\rho_{sv}\rho_{uw}$$

$$= \sum_{s,u=1}^{i}\{\sum_{v=1}^{i}|P_{(i)}|\rho^{su}_{(i)}\ \rho_{sv}\ \sum_{w=1}^{j}|P_{(j)}|\rho^{vw}_{(j)}\ \rho_{uw}\ +$$

$$\sum_{v=i+1}^{j}|P_{(i)}|\rho^{su}_{(i)}\ \rho_{sv}\ \sum_{w=1}|P_{(j)}|\rho^{vw}_{(j)}\ \rho_{uw}\}$$

$$= |P_{(j)}|(\sum_{s,u=1}^{i}|P_{(i)}|\rho^{su}_{(i)}\ \rho_{su}+0) = i\ |P_{(i)}||P_{(j)}|\ ,$$

$$(6.4)\qquad \sum_{s,u=1}^{i}\ \sum_{v,w=1}^{j}|P_{(i)}||P_{(j)}|\rho^{su}_{(i)}\rho^{vw}_{(j)}\ \rho_{su}\rho_{sv}\rho_{sw} =$$

$$|P_{(i)}|\ \sum_{s=1}^{i}\ \sum_{v,w=1}^{j}|P_{(j)}|\rho^{vw}_{(j)}\rho_{sv}\rho_{sw}$$

$$= |P_{(i)}|\ \sum_{s=1}^{i}\{\sum_{v=1}^{i}\ \rho_{sv}\ \sum_{w=1}^{j}|P_{(j)}|\rho^{vw}_{(j)}\ \rho_{sw}\ +\ \sum_{v=i+1}^{j}\ \rho_{sv}\ \sum_{w=1}^{j}|P_{(j)}|\rho^{vw}_{(j)}\rho_{sw}\}$$

$$= |P_{(i)}|\ (i|P_{(j)}|\ +\ 0) = i\ |P_{(i)}||P_{(j)}|\ ,$$

$$(6.5)\qquad \sum_{s,u=1}^{i}\ \sum_{v,w=1}^{j}|P_{(i)}||P_{(j)}|\rho^{su}_{(i)}\rho^{vw}_{(j)}\rho_{sv}\rho_{uv}\rho_{vw}$$

$$= |P_{(j)}|\ \sum_{s,u=1}^{i}\ \sum_{v=1}^{j}|P_{(i)}|\rho^{su}_{(i)}\ \rho_{sv}\rho_{uv}$$

$$= |P_{(j)}|\ \sum_{s,u=1}^{i}\{\sum_{v=1}^{i}\ |P_{(i)}|\rho^{su}_{(i)}\rho_{sv}\rho_{uv}\ +\ \sum_{v=i+1}^{j}|P_{(i)}|\rho^{su}_{(i)}\rho_{sv}\rho_{uv}\}$$

$$= i|P_{(i)}||P_{(j)}|+|P_{(i)}||P_{(j)}|\ \sum_{s,u=1}^{i}\rho^{su}_{(i)}\ \sum_{v=i+1}^{j}\rho_{sv}\rho_{uv}\ ,$$

we have from (6.1) that

$$\text{L-Cov}(|R_{(i)}|,|R_{(j)}|) = \frac{1}{n}[2|P_{(i)}||P_{(j)}| \sum_{s=1}^{i} \sum_{v=1}^{j} \rho_{sv}^{2} - 2i|P_{(i)}||P_{(j)}|$$

$$-2|P_{(i)}||P_{(j)}| \sum_{s,u=1}^{i} \rho_{(i)}^{su} a_{su}] ,$$

where $a_{su} = \sum_{v=i+1}^{j} \rho_{sv}\rho_{vu}$.

If we denote by P_{ij} the $i \times (j-i)$ matrix (ρ_{sv}), $s=1,\ldots,i;$ $v=i+1,\ldots,j,$ our result becomes

Theorem 6.1. The limiting joint distribution of

$$\sqrt{n}\ \{(|R_2|,\ldots,|R_k|) - (|P_2|,\ldots,|P_k|)\}$$

is normal with mean vector $(0,\ldots,0)$ and covariance matrix $H = (\eta_{ij})$, where

$$\eta_{ij} = 2|P_{(i)}||P_{(j)}|\{2 \sum_{s<u}^{i} \rho_{su}^{2} + \sum_{s=1}^{i} a_{ss} - \sum_{s,u=1}^{i} \rho_{(i)}^{su} a_{su}\}$$

$$= 2|P_{(i)}||P_{(j)}|\{\operatorname{tr} P_{(i)}^{2} - i + \operatorname{tr} P_{ij}'(I-P_{(i)}^{-1})\, P_{ij}\} .$$

7. Asymptotic joint distribution of determinants of a correlation matrix and of its principal submatrices.

Consider the following partition of the sample and the population correlation matrices:

$$R = \begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1q} \\ R_{21} & R_{22} & \cdots & R_{2q} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ R_{q1} & R_{q2} & \cdots & R_{qq} \end{bmatrix} , \quad P = \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1q} \\ P_{21} & P_{22} & \cdots & P_{2q} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ P_{q1} & P_{q2} & \cdots & P_{qq} \end{bmatrix} ,$$

where R_{ij} is of order $k_i \times k_j$, $i,j=1,\ldots,q$, and $k_1+\cdots+k_q = k$, and $k_i \geq 2$, $i=1,\ldots,q$.

In this section, we derive the asymptotic joint distribution of $|R|$, $|R_{11}|,\ldots,|R_{qq}|$. As before, the limiting joint distribution of

$$\sqrt{n}\ \{(|R|,|R_{11}|,\ldots,|R_{qq}|)-(|P|,|P_{11}|,\ldots,|P_{qq}|)\}$$

is normal with mean vector $(0,\ldots,0)$. From Theorem 5.1,

$$\text{(7.1)} \qquad \text{L-Var}|R_{ii}| = \frac{2}{n}|P_{ii}|^2(\text{tr } P_{ii}^2-k_i) , \quad i=0,1,\ldots,q ,$$

$(R_{oo} \equiv R)$ so that if $P_{ii} \neq I$, the variance is positive.

The covariance of any $|R_{ii}|$ and $|R|$ can be obtained as in the case of the step-down correlation matrices. If we define the $k_j \times (k-k_j)$ matrix

$$\overline{P}_j = (P_{j1},\ldots,P_{j(j-1)},P_{j(j+1)},\ldots,P_{jq}) ,$$

then

$$(7.2)\quad \text{L-Cov}(|R_{ii}|, |R|) = \frac{1}{n} 2|P_{ii}||P|[\operatorname{tr} P_{ii}^2 - k_i + \operatorname{tr}\{\bar{P}_i'(I_i - P_{ii}^{-1})\bar{P}_i\}] ,$$

where I_j is the $k_j \times k_j$ unit matrix.

Finally, we evaluate the covariance of $|R_{ii}|$ and $|R_{jj}|$, $(i \neq j)$, which cannot be obtained from the previous computations.

$$(7.3)\quad \text{L-Cov}(|R_{ii}|, |R_{jj}|) = \frac{1}{n} \sum_{s,u,v,w} |P_{ii}|\rho_{(ii)}^{su}|P_{jj}|\rho_{(jj)}^{vw} \times$$

$$\times \{\tfrac{1}{2} \rho_{su}\rho_{vw}(\rho_{sv}^2 + \rho_{sw}^2 + \rho_{uv}^2 + \rho_{uw}^2) + \rho_{sv}\rho_{uw} + \rho_{sw}\rho_{uv}$$

$$- (\rho_{su}\rho_{sv}\rho_{sw} + \rho_{su}\rho_{uv}\rho_{uw} + \rho_{sv}\rho_{uv}\rho_{vw} + \rho_{sw}\rho_{uw}\rho_{vw})\} ,$$

where $P_{ii}^{-1} = (\rho_{(ii)}^{su})$; the summation suffixes s,u run over $\{k_1 + \cdots + k_{i-1} + 1, \ldots, k_1 + \cdots + k_i\}$, the suffixes v,w run over $\{k_1 + \cdots + k_{j-1} + 1, \ldots, k_1 + \cdots + k_j\}$.

We now evaluate the individual terms of (7.3).

$$\sum_{s,u,v,w} |P_{ii}||P_{jj}|\rho_{(ii)}^{su}\rho_{(jj)}^{vw}\ \rho_{su}\rho_{vw}(\rho_{sv}^2 + \rho_{sw}^2 + \rho_{uv}^2 + \rho_{uw}^2)$$

$$= 4 \sum_{s,u,v,w} |P_{ii}||P_{jj}|\rho_{(ii)}^{su}\rho_{(jj)}^{vw}\ \rho_{su}\rho_{vw}\rho_{sv}^2$$

$$= 4 \sum_{s,v} \Big(\sum_u |P_{ii}|\rho_{(ii)}^{su}\ \rho_{su}\Big)\Big(\sum_w |P_{jj}|\rho_{(jj)}^{vw}\ \rho_{vw}\Big)\rho_{sv}^2$$

$$= 4\ |P_{ii}||P_{jj}| \sum_{s,v} \rho_{sv}^2 = 4|P_{ii}||P_{jj}|\ \operatorname{tr}(P_{ij}P_{ij}') ;$$

$$\sum_{s,u,v,w} |P_{ii}||P_{jj}|\rho^{su}_{(ii)}\rho^{vw}_{(jj)}(\rho_{sv}\rho_{uw} + \rho_{sw}\rho_{uv})$$

$$= 2 \sum_{s,u,v,w} |P_{ii}||P_{jj}|\rho^{su}_{(ii)}\rho^{vw}_{(jj)}\ \rho_{sv}\rho_{uw}$$

$$= 2|P_{ii}||P_{jj}| \sum_{s,u,v,w} \rho^{su}_{(ii)}\rho^{vw}_{(jj)}\ \rho_{sv}\rho_{uw}$$

$$= 2|P_{ii}||P_{jj}| \sum_{s,w} \Big(\sum_{u} \rho^{su}_{(ii)}\ \rho_{uw}\Big)\Big(\sum_{v} \rho^{vw}_{(jj)}\ \rho_{sv}\Big)$$

$$= 2|P_{ii}||P_{jj}|\ \mathrm{tr}[(P^{-1}_{ii}\ P_{ij})(P^{-1}_{jj}\ P'_{ij})] ,$$

$$\sum_{s,u,v,w} |P_{ii}||P_{jj}|\rho^{su}_{(ii)}\rho^{vw}_{(jj)}(\rho_{su}\rho_{sv}\rho_{sw}+\rho_{su}\rho_{uv}\rho_{uw}+\rho_{sv}\rho_{uv}\rho_{vw}+\rho_{sw}\rho_{uw}\rho_{vw})$$

$$= 2 \sum_{s,u,v,w} |P_{ii}||P_{jj}|\rho^{su}_{(ii)}\rho^{vw}_{(jj)}(\rho_{su}\rho_{sv}\rho_{sw} + \rho_{sv}\rho_{uv}\rho_{vw})$$

$$= 2\sum_{s} \Big(\sum_{u} |P_{ii}|\rho^{su}_{(ii)}\rho_{su}\Big) \sum_{v,w}|P_{jj}|\rho^{vw}_{(jj)}\rho_{sv}\rho_{sw} +$$

$$2 \sum_{s,u,v} |P_{ii}|\rho^{su}_{(ii)}\rho_{sv}\rho_{uv}\ \ \Big(\sum_{w} |P_{jj}|\rho^{vw}_{(jj)}\rho_{vw}\Big)$$

$$= 2|P_{ii}||P_{jj}|\Big\{ \sum_{v,w} \rho^{vw}_{(jj)}\ \Big(\sum_{s} \rho_{sv}\rho_{sw}\Big) + \sum_{s,u} \rho^{su}_{(ii)}\ \Big(\sum_{v} \rho_{sv}\rho_{uv}\Big)\Big\}$$

$$= 2|P_{ii}||P_{jj}|\ \mathrm{tr}(P_{ij}P^{-1}_{jj}\ P'_{ij} + P'_{ij}P^{-1}_{ii}P_{ij}) .$$

Combining these results, we obtain

$$(7.4)\quad \text{L-Cov}(|R_{ii}|,|R_{jj}|) = \frac{1}{n}\, 2|P_{ii}||P_{jj}|\{\mathrm{tr}(P_{ij}P'_{ij})+\mathrm{tr}(P_{ii}^{-1}P_{ij})(P_{jj}^{-1}P'_{ij})$$

$$- \mathrm{tr}(P_{ij}P_{jj}^{-1}P'_{ij} + P'_{ij}P_{ii}^{-1}P_{ij})\}$$

$$= \frac{1}{n}\, 2|P_{ii}||P_{jj}|\ \mathrm{tr}\{(P_{ii}^{-1}-I_i)P_{ij}(P_{jj}^{-1}-I_j)P'_{ij}\} .$$

If $i=j$ in (7.4), this covariance is reduced to the variance of $|R_{ii}|$.

If all P_{ij}, $i \neq j$, vanish, it is easily seen that the covariance matrix in the limiting distribution is singular. However, if at least one P_{ij} $(i \neq j)$ is not zero, then we have a non-singular limiting distribution.

Theorem 7.1. If $P_{ii} \neq I_i$, $i=1,\ldots,q$, and at least one of P_{ij} $(i \neq j)$ is not zero, then the limiting joint distribution of $\{(|R_{oo}|,\ldots,|R_{qq}|)\}$ is a $(q+1)$-variate (non-singular) normal distribution with mean vector $(|P_{oo}|,\ldots,|P_{qq}|)$ and covariance matrix $H = (\eta_{ij})$, where

$$\eta_{ij} = \mathrm{Cov}(|R_{ii}|,|R_{jj}|)$$

$$= \frac{1}{n}\, 2|P_{ii}||P_{jj}|\mathrm{tr}\{P_{ii}^{-1}-I_i)P_{ij}(P_{jj}^{-1}-I_j)P'_{ij}\}\ ,\ i,j=1,\ldots,q\ ,$$

$$(7.5)$$

$$\eta_{oi} = \text{L-Cov}(|R_{ii}|,|R|) = \frac{1}{n}\, 2|P_{ii}||P|\{\mathrm{tr}(P_{ii}^2)-k_i+\mathrm{tr}\ \overline{P}'_i(I_i-P_{ii}^{-1})\overline{P}_i\}$$

$$= \frac{1}{n}\, 2|P_{ii}||P| \sum_{j=1}^{q} \mathrm{tr}\{P'_{ij}(I_i-P_{ii}^{-1})P_{ij}\},\quad i=1,\ldots,q$$

$$\eta_{oo} = \text{L-Var}(|R|) = \frac{1}{n}\, 2|P|^2\{\mathrm{tr}(P^2)-k\} .$$

8. The non-central asymptotic distribution of the likelihood ratio criterion for testing the independence of sets of variates.

Let X be a random vector having a k-variate normal distribution. If we partition X into q subvectors with $k_1,\ldots,k_q$ components, i.e., $X = (X_1,\ldots,X_q)$, then the likelihood ratio criterion for testing the hypothesis that the subvectors are mutually independent, i.e., that $P_{ij} = 0$, for all $i \neq j$, is a monotone function of

$$(8.1) \qquad v = \frac{|S|}{|S_{11}| \cdots |S_{qq}|} = \frac{|R|}{|R_{11}| \cdots |R_{qq}|} .$$

Under the null hypothesis, the exact distribution of v is known for the special cases $q=2$, 3, and k. For the null case, a good approximation is given by $\mathcal{L}(-m \log v) = \mathcal{L}(\chi^2_d)$, where $d \equiv \frac{1}{2}(k^2-\Sigma k_i^2)$ degrees of freedom, and $m = n-\frac{1}{2}-\frac{1}{3}(k^3-\Sigma k_i^3)/(k^2-\Sigma k_i^2)$. However, the non-null distribution, i.e., the distribution when at least one of P_{ij}, $i \neq j$, is not the zero matrix is unknown.

Since v is a function of $R_{oo}, R_{11},\ldots,R_{qq}$, as a consequence of Theorems 2.1 and 7.1 we have that $\sqrt{n}\,\{v-\nu\}$, where $\nu = |P|/\prod_1^q |P_{ii}|$, is asymptotically normal with mean 0. From

$$\frac{\partial v}{\partial |P_{ii}|} = -\nu |P_{ii}|^{-1} \quad , \quad i=0,1,\ldots,q \ ;$$

we have that

$$(8.2) \qquad \text{L-Var}(v) = \nu^2 \sum_{i=1}^{q} \sum_{j=1}^{q} |P_{ii}|^{-1}|P_{jj}|^{-1} \text{ L-Cov}(|R_{ii}|, |R_{jj}|)$$

$$+ \nu^2|P|^{-2} \text{ L-Var}(|R|) - 2\nu^2 \sum_{i=1}^{q} |P|^{-1}|P_{ii}|^{-1}\text{L-Cov}(|R|,|R_{ii}|).$$

Using (7.5), these terms become

$$(8.3) \quad n\cdot\text{L-Var}(v) = 2\nu^2 \sum_{i=1}^{q} \sum_{j=1}^{q} \text{tr}\{(P_{ii}^{-1}-I_i)P_{ij}(P_{jj}^{-1}-I_j)P'_{ij}\}$$

$$+ 2\nu^2\{\text{tr}(P^2)-k\} - 4\nu^2 \sum_{i=1}^{q} \sum_{j=1}^{q} \text{tr}\{P'_{ij}(I_i-P_{ii}^{-1})P_{ij}\}$$

$$= 2\nu^2 \sum_{i=1}^{q} \sum_{j=1}^{q} \text{tr}\{P_{ii}^{-1}P_{ij}P_{jj}^{-1}P'_{ij}-P_{ij}P_{jj}^{-1}P'_{ij}+P_{ii}^{-1}P_{ij}P'_{ij}$$

$$- P_{ij}P'_{ij}\} + 2\nu^2\{\text{tr}(P^2)-k\} ,$$

and since

$$\sum_{i=1}^{q} \sum_{j=1}^{q} \text{tr}(P_{ij}P'_{ij}) = \text{tr}(P^2) ,$$

$$\sum_{i=1}^{q} \sum_{j=1}^{q} \text{tr}(P_{ij}P_{jj}^{-1}P'_{ij}) = \sum_{i=1}^{q} \sum_{j=1}^{q} \text{tr}(P_{ii}^{-1}P_{ij}P'_{ij}) ,$$

(8.3) becomes

$$(8.4) \qquad \text{L-Var}(v) = \frac{2\nu^2}{n} \left[\sum_{i=1}^{q} \sum_{j=1}^{q} \text{tr}\{P_{ii}^{-1}P_{ij}P_{jj}^{-1}P'_{ij}\}-k\right] .$$

A further simplification can be obtained by defining $D_P = \text{diag}(P_{11},\ldots,P_{qq})$, i.e., a diagonal matrix with blocks on the main diagonal, then

$$\sum_{i=1}^{q} \sum_{j=1}^{q} \operatorname{tr}(P_{ii}^{-1}P_{ij}P_{jj}^{-1}P_{ij}') = \operatorname{tr}(D_P^{-1}P)^2 ,$$

so that

$$\text{(8.5)} \qquad \text{L-Var}(v) = \frac{2\nu^2}{n} \{\operatorname{tr}(D_P^{-1}P)^2 - k\} .$$

To summarize, we have

Theorem 8.1. Let $v = |R|/\prod_1^q |R_{ii}|$, $\nu = |P|/\prod_1^q |P_{ii}|$, then $\sqrt{n}\,\{v-\nu\}$ is asymptotically normally distributed with mean 0 and with variance $\frac{2}{n}\nu^2[\operatorname{tr}(D_P^{-1}P)^2-k]$.

9. Non-central asymptotic distribution of the step-down multiple correlation coefficients.

Suppose that we have a k-variate (non-singular) normal population, $N(\mu,\Sigma)$, with correlation matrix P. In [2], Roy and Bargmann used the set of step-down multiple correlation coefficients for a test of independence. As in Section 6, let us use $R_{(i)}$ and $P_{(i)}$ as step-down sample and population correlation matrices, i.e., $R_{(\alpha)} = (r_{ij})$, $P_{(\alpha)} = (\rho_{ij})$, $i,j=1,\ldots,\alpha$.

Denote the sample and population multiple correlation coefficients between the i^{th} variate and the preceding variates by $\bar{r}_i$ and $\bar{\rho}_i$, respectively (the i^{th} step-down multiple correlation coefficient), which are abbreviated notations of $r_{i(1,\ldots,i-1)}$ and $\rho_{i(1,\ldots,i-1)}$. Then

$$(9.1)\qquad 1-\bar{r}_i^2 = \frac{|R_{(i)}|}{|R_{(i-1)}|}\ , \quad 1-\bar{\rho}_i^2 = \frac{|P_{(i)}|}{|P_{(i-1)}|}\ , \quad i=2,\ldots,k\ .$$

It is known that, under the null hypothesis of independence, i.e., when $P = I_k$, $(\bar{r}_1,\ldots,\bar{r}_k)$ are independently distributed.

Since we have already obtained the asymptotic joint distribution of $|R_{(1)}|,\ldots,|R_{(k)}|$, by making the transformations

$$|R_{(i)}| = (1-\bar{r}_2^2)\cdot\cdots\cdot(1-\bar{r}_i^2)\ , \quad i=2,\ldots,k\ ,$$

the Jacobian of which is

$$(1-\bar{r}_2^2)^{k-3}(1-\bar{r}_3^2)^{k-4}\cdot\cdots\cdot(1-\bar{r}_{k-1}^2)\ ,$$

we should be able to immediately obtain the desired result. Unfortunately, the result is of a complicated form, and consequently we again employ Theorem 2.1.

Define $f_i(R) = \bar{r}_i^2 = 1-|R_{(i)}|/R_{(i-1)}|$. Then $\sqrt{n}\ \{(\bar{r}_2^2,\ldots,\bar{r}_k^2)-(\bar{\rho}_2^2,\ldots,\bar{\rho}_k^2)\}$ is asymptotically normally distributed with mean vector 0 and with covariance matrix $\Phi = (\varphi_{ij})$. To evaluate φ_{ij}, we have

$$\frac{\varphi_{ii}}{n} = \text{L-Var}(\bar{r}_i^2) = \text{L-Var}(|R_{(i)}|/|R_{(i-1)}|)$$

$$= E\left\{ \left.\frac{\partial f_i(R)}{\partial |R_{(i)}|}\right|_{R=P} (|R_{(i)}|-|P_{(i)}|) + \left.\frac{\partial f_i(R)}{\partial |R_{(i-1)}|}\right|_{R=P} (|R_{(i-1)}|-|P_{(i-1)}|) \right\}^2$$

$$= (1-\bar{\rho}_i^2)^2 \left\{ \frac{1}{|P_{(i)}|^2} \text{L-Var}(|R_{(i)}|) + \frac{1}{|P_{(i-1)}|^2} \text{L-Var}(|R_{(i-1)}|) \right.$$

$$\left. - 2\frac{1}{|P_{(i)}||P_{(i-1)}|} \text{L-Cov}(|R_{(i)}|,|R_{(i-1)}|) \right\} ,$$

$$= \frac{2}{n}(1-\bar{\rho}_i^2)^2 [(\text{tr}P_i^2 - i) + (\text{tr}P_{i-1}^2 - i+1)$$

$$-2\{\text{tr}P_{i-1}^2 - i+1+a_i'(I_{i-1}-P_{i-1}^{-1})a_i\}] ,$$

where $a_i' = (\rho_{1i},\ldots,\rho_{(i-1)i})$.

Since

$$\text{tr}P_{i-1}^2 + 2a_i'a_i + 1 = \text{tr}P_i^2 ,$$

and

$$a_i'P_{i-1}^{-1}a_i = 1-\{1-a_i'P_{i-1}^{-1}a_i\} = 1 - \frac{|P_{(i)}|}{|P_{(i-1)}|} = \bar{\rho}_i^2 ,$$

we have

$$\varphi_{ii} = 4\,\bar{\rho}_i^2(1-\bar{\rho}_i^2)^2 \quad , \quad i=2,\ldots,k , \tag{9.2}$$

as to be expected.

Next we calculate:

$$\frac{\varphi_{ij}}{n} = \text{L-Cov}(\bar{r}_i^2, \bar{r}_j^2) = \text{L-Cov}\left(\frac{|R_{(i)}|}{|R_{(i-1)}|} \;,\; \frac{|R_{(j)}|}{|R_{(j-1)}|}\right) \qquad (i < j)$$

$$= (1-\bar{\rho}_i^2)(1-\bar{\rho}_j^2)\left\{ \frac{1}{|P_{(i)}||P_{(j)}|} \text{L-Cov}(|R_{(i)}|,|R_{(j)}|) \right.$$

$$+ \frac{1}{|P_{(i-1)}||P_{(j-1)}} \text{L-Cov}(|R_{(i-1)}| \;,\; |R_{(j-1)}|)$$

$$- \frac{1}{|P_{(i)}||P_{(j-1)}|} \text{L-Cov}(|R_{(i)}|,|R_{(j-1)}|)$$

$$\left. - \frac{1}{|P_{(i-1)}||P_{(j)}|} \text{L-Cov}(|R_{(i-1)}|,|R_{(j)}|) \right\}$$

$$= \frac{2}{n}(1-\bar{\rho}_i^2)(1-\bar{\rho}_j^2)[\{\text{tr}P_i^2 - i + \text{tr}(P'_{ij}(I_i - P_i^{-1})P_{ij})\}$$

$$+ \{\text{tr}P_{i-1}^2 - i + 1 + \text{tr}P'_{i-1,j-1}(I_{i-1} - P_{i-1}^{-1})P_{i-1,j-1}\}$$

$$- \{\text{tr}P_i^2 - i + \text{tr}P'_{i,j-1}(I_i - P_i^{-1})P_{i,j-1}\}$$

$$- \{\text{tr}P_{i-1}^2 - i + 1 + \text{tr}P'_{i-1,j}(I_{i-1} - P_{i-1}^{-1})P_{i-1,j}\}]$$

$$= \frac{2}{n}(1-\rho_i^{-2})(1-\rho_j^{-2})\{\text{tr}P'_{ij}(I_i - P_i^{-1})P_{ij} + \text{tr}P'_{i-1,j-1}(I_{i-1} - P_{i-1}^{-1})P_{i-1,j-1}$$

$$-\text{tr}P'_{i,j-1}(I_i - P_i^{-1})P_{i,j-1} - \text{tr}P'_{i-1,j}(I_{i-1} - P_{i-1}^{-1})P_{i-1,j}\} ,$$

where P_{ij} is an $i \times (j-i)$ matrix consisting of elements ρ_{sv}, $s=1,\dots,i$; $j=i+1,\dots,j$.

$$\mathrm{tr}P'_{ij}P_{ij} + \mathrm{tr}P'_{i-1,j-1}P_{i-1,j-1} - \mathrm{tr}P'_{i,j-1}P_{i,j-1} - \mathrm{tr}P'_{i-1,j}P_{i-1,j} = \rho^2_{ij},$$

and if we use the notation

$$b_j = (\rho_{1j},\dots,\rho_{ij}) \quad ; \; c_j = (\rho_{1j},\dots,\rho_{i-1,j}) \; ,$$

then

$$\begin{aligned} &\mathrm{tr}P'_{ij}P_i^{-1}P_{ij} - \mathrm{tr}P'_{i,j-1}P_i^{-1}P_{i,j-1} \\ &\quad = \mathrm{tr}P_i^{-1}\left([P_{i,j-1}\, b'_j] \begin{bmatrix} P'_{i,j-1} \\ \cdots \\ b_j \end{bmatrix} - P_{i,j-1}P'_{i,j-1}\right) \\ &\quad = \mathrm{tr}P_i^{-1}b'_jb_j = b_jP_i^{-1}b'_j \; , \end{aligned}$$

and similarly

$$\mathrm{tr}P'_{i-1,j}P^{-1}_{i-1}P_{i-1,j} - \mathrm{tr}P'_{i-1,j-1}P^{-1}_{i-1}P_{i-1,j-1} = c_jP^{-1}_{i-1}c'_j \; .$$

Hence

$$(9.3) \qquad \varphi_{ij} = 2(1-\bar{\rho}_i^2)(1-\bar{\rho}_j^2)(\rho^2_{ij} - b'_jP_i^{-1}b_j + c'_jP^{-1}_{i-1}c_j) \; .$$

In summary, we have

Theorem 9.1. If none of the step-down population multiple correlation coefficients, $\bar{\rho}_i^2$, $i=2,\dots,k$, are zero, then the step-down sample correlation coefficients, $\bar{r}_i^2$, $i=2,\dots,k$, have an asymptotic joint distribution which is $(k-1)$-variate normal with means, $\bar{\rho}_i^2$, $i=2,\dots,k$, and with covariance matrix, Φ/n, where φ_{ij} is given by (9.2) and (9.3).

REFERENCES

[1] Anderson, T.W., (1958). Introduction to Multivariate Statistical Analysis. Wiley, New York.

[2] Roy, S.N., and Bargmann, R.E., (1958). "Tests of multiple independence and the associated confidence bounds," Ann. Math. Statist., 29, 491-503.

ORDER STATISTICS IN TWO-DIMENSIONS

by

Minoru Siotani

0. Introduction.

Although there are many results and tables for order statistics, only a few deal with the multidimensional case. Recently, the author began to investigate the properties of order statistics defined in Section 1 in the two dimensional case. This report is not a presentation of the systematic theory of bivariate order statistics. It presents partial results recently obtained by the author. In Section 1, two-dimensional order statistics are defined; certain distribution results are obtained in Section 2. Problems concerning joint distributions are covered in Section 3. Section 4 deals with the bivariate range, and Section 5 with quantiles.

1. Two-dimensional order statistics (Definition).

Let $(x_1^*,y_1^*), (x_2^*,y_2^*),\dots,(x_n^*,y_n^*)$ be n two-dimensional random vectors. We rearrange n values of each component, separately, in ascending order of magnitude:

$$x_1 \leqq x_2 \leqq \cdots \leqq x_n ,$$
$$y_1 \leqq y_2 \leqq \cdots \leqq y_n .$$

Definition 1. The combined quantities (x_i,y_j), $i,j=1,\dots,n$ are called the $(i,j)^{th}$ two-dimensional or bivariate order statistics, and x_i, $i=1,\dots,n$, and y_j, $j=1,\dots,n$, are called, respectively, the i^{th} x-component and the j^{th} y-component order statistics.

Throughout this report, we assume that the underlying population is of continuous type with a cumulative distribution function (c.d.f.) $F(x,y)$ and a density function (d.f.) $f(x,y)$ defined over two-dimensional Euclidean space, R_2. So, $f(x,y)$ is assumed to be continuous and finite almost everywhere in R_2 with respect to the ordinary Lebesgue measure. If a particular distribution is defined over a finite domain, for example, $a < x \leqq b$; $c < y \leqq d$, it should be understood that $f(x,y)$ is zero outside of this domain.

2. The density functions of individual bivariate order statistics.

2.1. The $(n,n)^{th}$ order statistic.

The c.d.f. of the $(n,n)^{th}$ order statistic, (x_n,y_n), is obviously $F^n(x_n,y_n)$, and hence, by differentiation, the d.f. is

$$\begin{aligned}\phi_n(x_n,y_n) &= n[F(x_n,y_n)]^{n-1}f(x_n,y_n) \\ &\quad + n(n-1)[F(x_n,y_n)]^{n-2}F(x_n\|y_n)F(y_n\|x_n),\end{aligned} \tag{2.1}$$

where

$$F(x_n\|y_n) = \int_{-\infty}^{x_n} f(x,y_n)dx \quad \text{, and} \quad F(y_n\|x_n) = \int_{-\infty}^{y_n} f(x_n,y)dy \ . \tag{2.2}$$

For the derivation of exact d.f.'s of various bivariate order statistics, it is very useful to notice the meaning of each term of the right hand side in (2.1). The first term corresponds to the case that the $(n,n)^{th}$ order statistic is determined by single sample point and the second term corresponds to the case that the $(n,n)^{th}$ order statistic is determined by two different sample points (see Figure 1.).

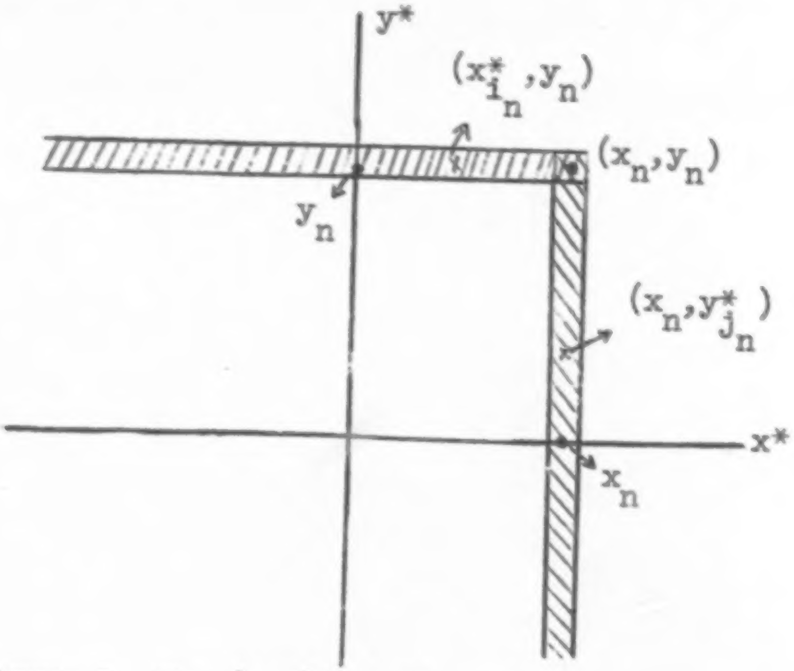


Figure 1. The (n,n)th order statistic.

By a similar consideration, we can easily obtain the exact d.f.'s of the (n,1)th, the (1,n)th and the (1,1)th order statistics. Let us summarize the results below:

2.2. The (n,1)th order statistic.

$$(2.3)\quad \begin{aligned}\phi_n(x_n,y_1) &= n[F_1(x_n)-F(x_n,y_1)]^{n-1}f(x_n,y_1)\\ &\quad + n(n-1)[F_1(x_n)-F(x_n,y_1)]^{n-2}F(x_n\|y_1)\{f_1(x_n)-F(y_1\|x_n)\}\end{aligned}$$

where $F_1(\cdots)$ and $f_1(\cdots)$ are, respectively, c.d.f. and d.f. of the marginal distribution of x.

2.3. The (1,n)th order statistic.

$$(2.4)\quad \begin{aligned}\phi_n(x_1,y_n) &= n[F_2(y_n)-F(x_1,y_n)]^{n-1}f(x_1,y_n)\\ &\quad + n(n-1)[F_2(y_n)-F(x_1,y_n)]^{n-2}F(y_n\|x_1)\{f_2(y)-F(x_1\|y_n)\},\end{aligned}$$

where $F_2(\cdots)$ and $f_2(\cdots)$ are, respectively, c.d.f. and d.f. of the marginal distribution of y.

2.4. The (1,1)th order statistic.

$$(2.5)\quad \begin{aligned}\phi_n(x_1,y_1) &= n[1-F_1(x_1)-F_2(y_1)+F(x_1,y_1)]^{n-1}f(x_1,y_1)\\ &\quad + n(n-1)[1-F_1(x_1)-F_2(y_1)+F(x_1,y_1)]^{n-2}\{f_1(x_1)-F(y_1\|x_1)\}\cdot\\ &\quad \cdot\{f_2(y_1)-F(x_1\|y_1)\}\end{aligned}$$

2.5. The (i,j)th order statistic.

The d.f. of the (i,j)th order statistic has a more complicated form for i and j different from n and 1. However, the principle of the derivation is similar as seen below.

Let i and j be any intermediate integers between n and 1. There are two cases to be considered in the formation of the $(i,j)^{th}$ order statistic (x_i,y_j), i.e.,

(a) x_i and y_j are components of single sample point,

(b) x_i and y_j come from two different sample points, say, (x^*_α,y^*_α) with $x^*_\alpha = x_i$ and $y^*_\alpha \neq y_j$ and (x^*_β,y^*_β) with $x^*_\beta \neq x_i$ and $y^*_\beta = y_j$.

The possibility that there are two or more sample points having the same x_i or y_j may be omitted, because we are concerning the continuous case, and the probability of occurrence of such event is zero. In the present intermediate cases, case (b) is realized in one of the following four cases:

(b.1) $x^*_\beta < x_i$, $y^*_\alpha < y_j$, (b.2) $x^*_\beta < x_i$, $y^*_\alpha > y_j$,

(b.3) $x^*_\beta > x_i$, $y^*_\alpha < y_j$, (b.4) $x^*_\beta > x_i$, $y^*_\alpha > y_j$,

(Figure 2). Since the five cases (a), (b.1), (b.2), (b.3) and (b.4) are mutually exclusive, the d.f. of (x_i,y_j) can be derived by adding the probability elements for these cases. To do so, we use the notation:

R_o: $\{(x^*,y^*):\ x_i - \frac{1}{2}dx_i \leqq x^* < x_i + \frac{1}{2}dx_i,\ y_j - \frac{1}{2}dy_j \leqq y^* < y_j + \frac{1}{2}dy_j\}$,

R_{11}: $\{(x^*,y^*):\ -\infty < x^* < x_i - \frac{1}{2}dx_i\ ,\ -\infty < y^* < y_j - \frac{1}{2}dy_j\}$,

R_{12}: $\{(x^*,y^*):\ -\infty < x^* < x_i - \frac{1}{2}dx_i\ ,\ y_j + \frac{1}{2}dy_j \leqq y^* < \infty\}$,

R_{21}: $\{(x^*,y^*):\ x_i + \frac{1}{2}dx_i \leqq x^* < \infty\ ,\ -\infty < y^* < y_j - \frac{1}{2}dy_j\}$,

R_{22}: $\{(x^*,y^*):\ x_i + \frac{1}{2}dx_i \leqq x^* < \infty\ ,\ y_j + \frac{1}{2}dy_j \leqq y^* < \infty\}$,

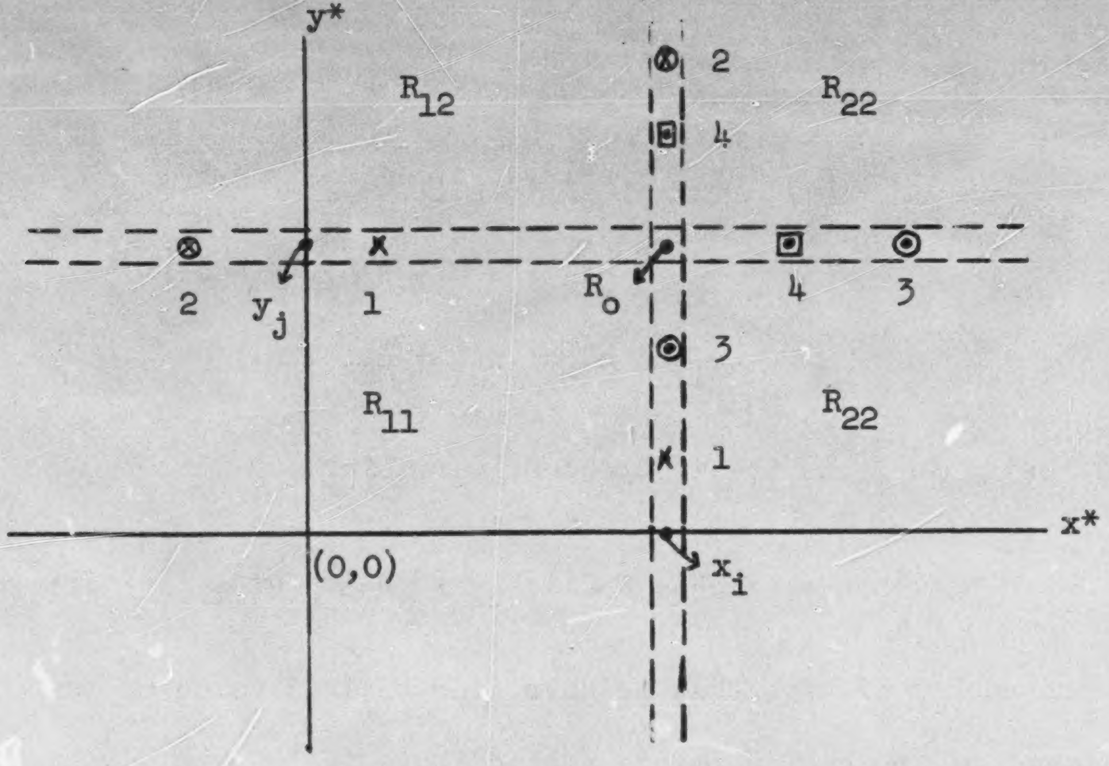


Figure 2. The order statistic (x_i,y_j). Four cases for the location of (x_i,y^*_α) and (x^*_β,y_j). The numbers i=1,2,3,4 stand for two points corresponding to cases (b.i).

n_{ij}: the number of sample points falling in the region R_{ij},

$P_{11} = F(x_i,y_j)$, $P_{12} = F_1(x_i)-F(x_i,y_j)$,

$P_{21} = F_2(y_j)-F(x_i,y_j)$, $P_{22} = 1-F_1(x_i)-F_2(y_j)+F(x_i,y_j)$.

Let us, first, consider the case (a). The probability element of (x_i,y_j) can be obtained by calculating the probability of the event that among n sample points, one is in R_o, n_{11} points are in R_{11}, n_{12} points are in R_{12}, n_{21} points in R_{21}, and n_{22} points are in R_{22}, and then neglecting the terms of smaller order than $dx_i dy_j$. To do so, it should be noted that, for fixed i and j, n_{ij} are subject to the restrictions,

$$\begin{gathered} n_{11} + n_{12} = i-1 , \quad n_{11} + n_{21} = j-1 , \\ n_{21} + n_{22} = n-i , \quad n_{12} + n_{22} = n-j , \\ n_{11} + n_{12} + n_{21} + n_{22} = n-1 \end{gathered} \tag{2.6}$$

and so, using n_{11} as the independent variable,

$$n_{12} = i-1-n_{11} , \quad n_{21} = j-1-n_{11} , \quad n_{22} = n+n_{11}+1-i-j . \tag{2.7}$$

Since the number of ways that we have, for a fixed value of n_{11}, the arrangement of n sample points stated above is

$\dfrac{n!}{|!n_{11}!(i-1-n_{11})!(j-1-n_{11})!(n+1-i-j+n_{11})!|}$ and since n_{11} assumes one of possible values, $0,1,\ldots,m-1$, where $m = \min(i,j)$, it is easy to see that the probability element for case (a) can be written as

$$\sum_{n_{11}=o}^{m-1} C(1,n_{11},i-1-n_{11},j-1-n_{11};n) P_{11}^{n_{11}} P_{21}^{j-1-n_{11}} P_{22}^{n+n_{11}+1-i-j} \cdot f(x_i,y_j)dx_i dy_j , \tag{2.8}$$

where

$$C(a,b,d,c;n) = \frac{n!}{a!b!c!d!(n-a-b-c-d)!} , \tag{2.9}$$

where $C(\cdots)$ is zero if any of the arguments is negative.

For each case (b.i), $i=1,2,3,4$, we can derive the (conditional) probability element in the same manner by keeping in mind the positions of two points determining the order statistic (x_i, y_j). We easily have the following results:

For the case (b.1),

$$\sum_{n_{11}=0}^{m-2} C(1,1,n_{11},i-2-n_{11},j-2-n_{11};n)P_{11}^{n_{11}}P_{12}^{i-2-n_{11}}P_{21}^{j-2-n_{11}}P_{22}^{n+n_{11}+2-i-j} \cdot F(x_i\|y_j)F(y_j\|x_i)dx_i dy_j \tag{2.10}$$

for the case (b.2),

$$\sum_{n_{11}=0}^{m^*} C(1,1,n_{11},i-2-n_{11},j-1-n_{11};n)P_{11}^{n_{11}}P_{12}^{i-2-n_{11}}P_{21}^{j-1-n_{11}}P_{22}^{n+n_{11}+1-i-j} \cdot F(x_i\|y_j)\{f_1(x_i)-F(y_j\|x_i)\}\cdot \cdot dx_i dy_j\ , \tag{2.11}$$

where $m^* = \min(i-2,j-1)$,

for the case (b.3),

$$\sum_{n_{11}=0}^{m^{**}} C(1,1,n_{11},i-1-n_{11},j-2-n_{11},n)P_{11}^{n_{11}}P_{12}^{i-1-n_{11}}P_{21}^{j-2-n_{11}}P_{22}^{n+n_{11}+1-i-j} \cdot F(y_j\|x_i)\{f_2(y_j)-F(x_i\|y_j)\}\cdot \cdot dx_i dy_j\ , \tag{2.12}$$

where $m^{**} = \min(i-1,j-2)$, and finally for the case (b.4),

$$\sum_{n_{11}=o}^{m-1} C(1,1,n_{11},i-1-n_{11},k-1-n_{11};n)P_{11}^{n_{11}}P_{12}^{i-1-n_{11}}P_{21}^{j-1-n_{11}}P_{22}^{n+n_{11}-i-j}$$

$$(2.13) \qquad \cdot\{f_1(x_i)-F(y_j\|x_i)\}\{f_2(y_j)-F(x_i\|y_j)\}\cdot$$

$$\cdot dx_i dy_j \ .$$

Thus the full probability element of (x_i,y_j) is given by the sum of (2.8), (2.10), (2.11), (2.12), and (2.13). So far, n and 1 are excluded from values of i and j. However, if we keep in mind the convention made on the coefficients, $C(\cdots)$, then it is easy to see that these extreme cases are also included in the above result. For example, for the case when $i=j=n$, all terms in (2.11), (2.12) and (2.13) are zero, and in (2.8) and (2.10), only the terms for $n_{11} = m-1=n-1$ and for $n_{11}=m-2=n-2$ remain respectively, the sum of which coincides with the density function in (2.1) times $dx_n dy_n$. Let us summarize the above in the

Theorem 2.1. The probability element of the $(i,j)^{th}$ order statistic, (x_i,y_j), $i,j=1,2,\ldots,n$, is

$$(2.14) \qquad \phi_n(x_i,y_j)dx_i dy_j = (2.8)+(2.10)+(2.11)+(2.12)+(2.13).$$

2.6. The two dimensional or the bivariate median.

Let us assume, for simplicity, that the sample size n is odd, say, $n=2s+1$.

Definition 2. The $(s+1,s+1)$th order statistic, (x_{s+1},y_{s+1}) is called a two-dimensional or a bivariate median in a sample of size $n=2s+1$.

The density function of the bivariate median is, of course, given by Theorem 2.1 with $n=2s+1$, $i=j=s+1$. This special case was treated by Mood [4] and Siddiqui [6]. We here present an alternative form of the d.f.

Theorem 2.2. *The density function of the bivariate median*, $(\tilde{x},\tilde{y}) \equiv (x_{s+1},y_{s+1})$, *in a sample of size* $n=2s+1$ *is given by the formula*,

$$\phi_n(\tilde{x},\tilde{y}) = \phi_n^{(1)}(\tilde{x})\cdot\phi_n^{(2)}(\tilde{y}|\tilde{x}) , \tag{2.15}$$

where

$$\phi_n^{(1)}(\tilde{x}) = \frac{(2s+1)!}{s!s!1!}\,[F_1(\tilde{x})]^s[1-F_1(\tilde{x})]^s f_1(\tilde{x}) , \tag{2.16}$$

and

$$\begin{aligned}\phi_n^{(2)}(\tilde{y}|\tilde{x}) = {} & \sum_{r=o}^{s} b(r;s,P_2)b(r;s,1-Q_2)f_2(\tilde{y}|\tilde{x}) \\ & + s^2\sum_{r=o}^{s-1} b(r;s-1,P_2)b(r;s-1,1-Q_2)\times \\ & \times \Big[\frac{1}{r+1}\{P_2q_2(1-F_2(\tilde{y}|\tilde{x})) + (1-Q_2)p_2F_2(\tilde{y}|\tilde{x})\} \\ & + \frac{1}{s-r}\{Q_2p_2(1-F_2(\tilde{y}|\tilde{x})) + (1-P_2)q_2F_2(\tilde{y}|\tilde{x})\}\Big] ,\end{aligned} \tag{2.17}$$

where $b(r;N,P) = \binom{N}{r}P^r(1-P)^{N-r}$, $f_2(\tilde{y}|\tilde{x}) = f(\tilde{x},\tilde{y})/f_1(\tilde{x})$,

$$F_2(\tilde{y}|\tilde{x}) = F(\tilde{y}\|\tilde{x})/f_1(\tilde{x}) ,\quad P_2 = F(\tilde{x},\tilde{y})/F_1(\tilde{x}),\quad Q_2 = \{F_2(\tilde{y})-F(\tilde{x},\tilde{y})\}/\{1-F_1(\tilde{x})\},$$

$$p_2 = \frac{\partial P_2}{\partial\tilde{y}} = F(\tilde{x}\|\tilde{y})/F_1(\tilde{x}) ,\quad q_2 = \frac{\partial Q_2}{\partial\tilde{y}} = \{f_2(\tilde{y})-F(\tilde{x}\|\tilde{y})\}/\{1-F_1(\tilde{x})\} .$$

3. Some joint density functions of bivariate order statistics.

3.1. The density function of the (n,n)th and the (1.1)th order statistic.

The density function can be easily obtained by geometrical considerations similar to the case for the individual order statistics, that is, by computing the probability that (x_1,y_1) will lie in the infinitesimal rectangle: $x_1 - \frac{1}{2} x_1 \leqq x^* < x_1 + \frac{1}{2} dx_1$, $y_1 - \frac{1}{2} dy_1 \leqq y^* < y_1 + \frac{1}{2} dy_1$ and (x_n,y_n) will lie in the infinitesimal rectangle: $x_n - \frac{1}{2} dx_n \leqq x^* < x_n + \frac{1}{2} dx_n$, $y_n - \frac{1}{2} dy_n \leqq y^* < y_n + \frac{1}{2} dy_n$.

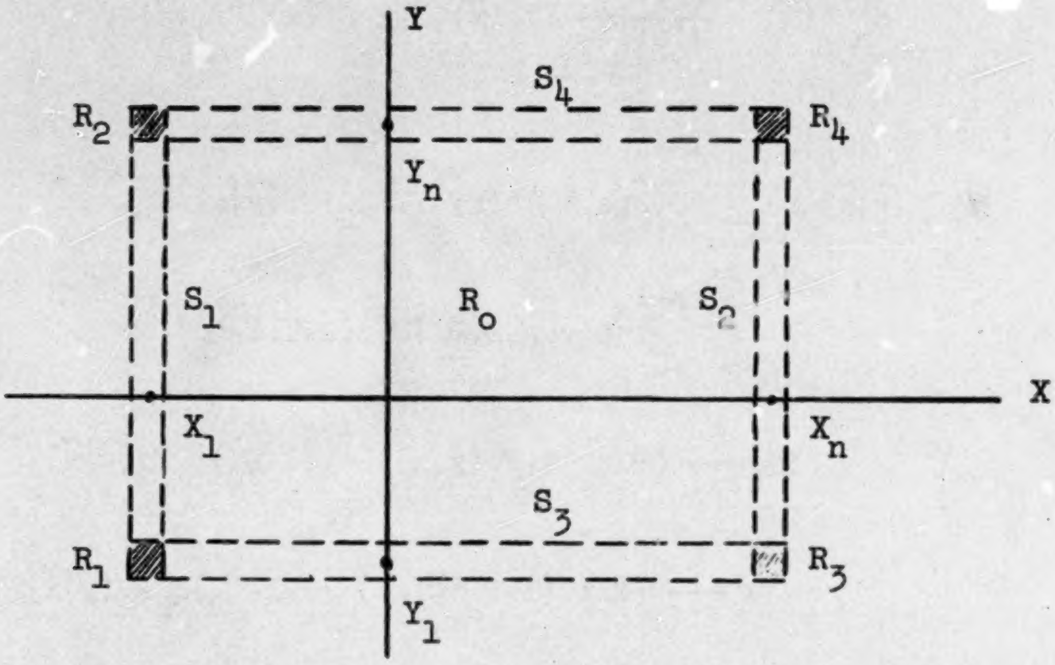


Figure 3. The order statistics (x_1,y_1) and (x_n,y_n).

With the reference of Figure 3, it is easy to see that the above event will be realized in one and only one of the following seven mutually exclusive cases:

(a) (x_1,y_1) and (x_n,y_n) are, respectively, sample points,

(b) Both of (x_1,y_n) and (x_n,y_1) are sample points,

(c) Both order statistics are determined from the three different sample points, in which there are four mutually exclusive possibilities:

(c.1) (x_1,y_1) is a sample point and (x_n,y_n) is determined by two points, one is in S_2 and the other is in S_4.

(c.2) (x_1,y_n) is a sample point and (x_n,y_1) is determined by two points lying in strips, S_2 and S_3 respectively,

(c.3) (x_n,y_1) is a sample point and (x_1,y_n) is determined by two points lying in strips, S_1 and S_4, respectively,

(c.4) (x_n,y_n) is a sample point and (x_1,y_1) is determined by two points lying in strips, S_1 and S_3, respectively.

(d) Both order statistics are determined from the four different sample points falling in strips, S_1, S_2 S_3 and S_4, respectively.

Then, computing the probability element for each case and summing, we obtain the desired joint density function multiplied by $dx_1 dx_n dy_1 dy_n$. We use the notation

$$P = F(x_n,y_n)-F(x_n,y_1)-F(x_1,y_n)+F(x_1,y_1) \ , \quad f(x_i,y_j) = f(i,j) \ .$$

$$p_j \equiv F(x_n \| y_j)-F(x_1 \| y_j) = \int_{x_1}^{x_n} f(x,y_j)dx \ , \quad j=1,n \ ,$$

$$q_i \equiv F(y_n \| x_i)-F(y_1 \| x_i) = \int_{y_1}^{y_n} f(x_i,y)dy \ , \quad i=1,n \ .$$

The result obtained in this way is summarized in the

Theorem 3.1. The joint density function of the (1,1)th and the (n,n)th order statistics is given by

$$\begin{aligned} \varphi_n(x_1,y_1,x_n,y_n) &= n(n-1)P^{n-2}\{f(1,1)f(n,n) + f(1,n)f(n,1)\} \\ &+ n(n-1)(n-2)P^{n-3}\{p_n q_n f(1,1) + p_1 q_n f(1,n) + p_n q_1 f(n,1) \\ &+ p_1 q_1 f(n,n)\} + n(n-1)(n-2)(n-3)P^{n-4}\, p_1 q_1 p_n q_n \ . \end{aligned} \tag{3.1}$$

If x^* and y^* are independent, (3.1) is, of course, reduced to the product of marginal density functions of x_1, x_n and y_1, y_n, i.e.,

$$\varphi_n(x_1,y_1,x_n,y_n) = [n(n-1)\{F_1(x_n)-F_1(x_1)\}^{n-2}f_1(x_n)f_1(x_1)]$$
$$\cdot[n(n-1)\{F_2(y_n)-F_2(y_1)\}^{n-2}f_2(y_1)f_2(y_n)].$$

Remark 1. For the present extreme case, Theorem 3.1 can be derived simply by differentiating the probability function

$$(3.2) \qquad P\{x_1 \leqq x^* \leqq x_n,\ y_1 \leqq y^* \leqq y_n\} = [F(x_n,y_n)-F(x_n,y_1)-F(x_1,y_n) + F(x_1,y_1)]^n$$

with respect to x_1, y_n, y_1 and y_n, with the rule we take the minus derivatives with respect to x_1 and y_1, because the probability function is a decreasing function of x_1 and y_1. But the above geometrical consideration also serves for other more complicated cases.

Remark 2. It is easily seen that the phrase "the joint d.f. of the (1,1)th and (n,n)th order statistics" is equivalent to the phrase "the joint d.f. of the (1,n)th and the (n,1)th order statistics". In fact, this is the joint d.f. of x_1, x_n, y_1 and y_n.

3.2. The joint density function of the (h,j)th and the (i,k)th order statistics.

As an example of more complicated cases, consider the joint density function of the (h,j)th and the (i,k)th order statistics. By remark 2 in Section 3.1, this is the joint d.f. of x_h, x_i, y_j and y_k, where $h < i$ and $j < k$. As before, the geometrical interpretation is useful (see Figure 4). We use the notation:

$$P_{\alpha\beta} = \int\int_{R_{\alpha\beta}} f(x,y)dxdy \ , \quad p_{\alpha\beta} = \int_{r_{\alpha\beta}} f(x,y_\beta)dx \ , \quad q_{\alpha\beta} = \int_{S_{\alpha\beta}} f(x_\alpha,y)dx$$

$$\alpha,\beta = 1,2,3 \qquad\qquad \alpha=1,2,3;\ \beta = 1,2 \qquad\qquad \alpha=1,2;\ \beta=1,2,3$$

$n_{\alpha\beta}$ = the number of sample points falling in the region $R_{\alpha\beta}$.

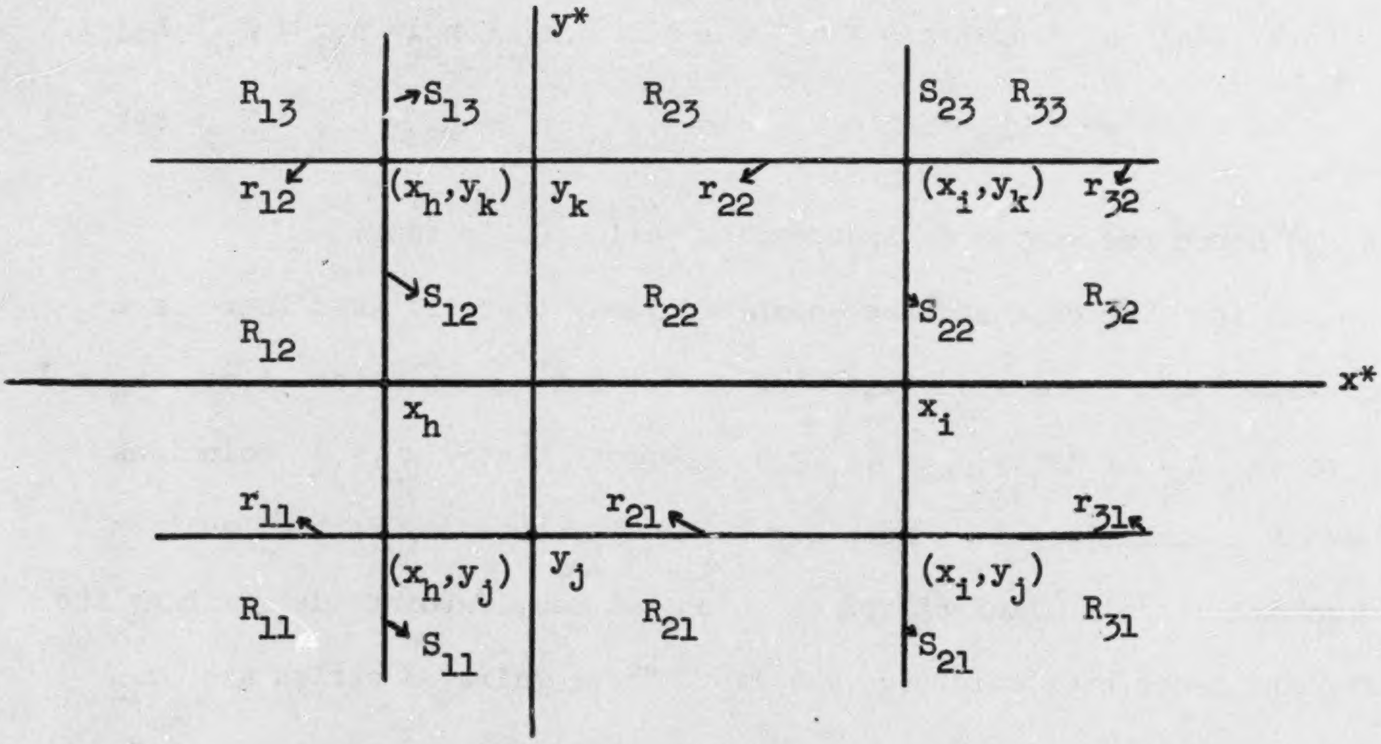

y*
R13
S13
R23
S23
R33
r12
(xh,yk)
yk
r22
(xi,yk)
r32
S12
R22
S22
R32
R12
x*
xh
xi
r11
r21
r31
(xh,yj)
yj
(xi,yj)
R11
S11
R21
S21
R31

Figure 4. The order statistics (x_h,y_j), (x_i,y_k), (x_h,y_k) and (x_i,y_j).

Let us consider the case by case according to the number of sample points determining our order statistics.

(a) The case two sample points, that is,

(a.1) when both of the (h,j)th and the (i,k)th order statistics are sample points or (a.2) when both of the (h,k)th and the (i,j)th order statistics are sample points.

$$(3.3) \qquad \{f(h,j)f(i,k) + f(h,k)f(i,j)\} \left[\sum_{L(n)} \frac{n!}{\prod_{\alpha,\beta} n_{\alpha\beta}!} \prod_{\alpha,\beta} P_{\alpha\beta}^{n_{\alpha\beta}} \right]$$

where the summation is carried out under the restrictions

$$(3.4) \quad L(n): \begin{array}{ll} n_{11}+ n_{12}+ n_{13} \equiv N_{1.} = h-1, & n_{11}+ n_{21}+ n_{31} \equiv N_{.1}= j-1, \\ n_{21}+ n_{22}+ n_{23} \equiv N_{2.} = i-h-1, & n_{12}+ n_{22}+ n_{32} \equiv N_{.2}= k-j-1, \\ n_{31}+ n_{32}+ n_{33} \equiv N_{3.} = n-i, & n_{13}+ n_{23}+ n_{33} \equiv N_{.3}= n-k, \end{array}$$

and hence the number of independent variables is four.

(b) The case of three sample points, that is, when there is a sample point consisting with one of our order statistics, i.e., (x_h,y_j) or (x_h,y_k) or (x_i,y_j) or (x_i,y_k). Suppose that (x_h,y_j) coincides with a sample point. There are nine ways according to the location of the infinitesimal strips on which two sample points determining the other order statistic (x_i,y_k) lie. These pairs of strips are: r_{12}, S_{21}; r_{12}, S_{22}; r_{12}, S_{23}; r_{22}, S_{21}; $r_{\beta\beta}$, S_{22}; $r_{\beta\beta}$, S_{23}; r_{32}, S_{23}; r_{32}, S_{22}; r_{32}, S_{23}. For this minor case, the contribution to the whole d.f. can be written as

$$(3.5) \qquad f(h,j) \sum_{a=1}^{3} \sum_{b=1}^{3} p_{a2}\, q_{2b} \left[\sum_{L_{a,b}(n)} \frac{n!}{\prod_{\alpha,\beta} n_{\alpha\beta}!} \prod_{\alpha,\beta} P_{\alpha\beta}^{n_{\alpha\beta}} \right],$$

where the summation in the brackets is taken over the values of $n_{\alpha\beta}$'s satisfying the restrictions,

$$(3.6)\quad L_{a,b}(n):\ \begin{array}{l} N_{1.} = h-1-\delta_{1a},\ \ N_{2.} = i-h-1-\delta_{2a},\ \ N_{3.} = n-i-\delta_{3a} \\ N_{.1} = j-1-\delta_{1b},\ \ N_{.2} = k-j-1-\delta_{2b},\ \ N_{.3} = n-k-\delta_{3b} \end{array}$$

where δ_{1a}, $\delta_{1b}, \ldots$ are Kronecker's symbols.

Other minor cases in the case (b) are evaluated in the same way. Summing up these results, we can then have the contribution due to the case (b) in the form

$$(3.7)\quad \sum_{a=1}^{3} \sum_{b=1}^{3} \{f(h,j)p_{a2}q_{2b}+f(h,k)p_{a1}q_{2b}+f(i,j)p_{a2}q_{1b}+f(i,k)p_{a1}q_{1b}\} \cdot \left[\sum_{L_{a,b}(n)} \frac{n!}{\prod_{\alpha,\beta} n_{\alpha\beta}!} \prod_{\alpha,b} p_{\alpha\beta}^{n_{\alpha\beta}} \right].$$

(c) The case of four points, that is, when each of (x_h, y_j) and (x_i, y_k) or each of (x_h, y_k) and (x_i, y_j) are determined from two sample points. In a similar manner, we have

$$(3.8)\quad \sum_{a=1}^{3} \sum_{b=1}^{3} \sum_{c=1}^{3} \sum_{d=1}^{3} p_{a1}p_{b2}q_{1c}q_{2d} \left[\sum_{L_{a,b,c,d}(n)} \frac{n!}{\prod_{\alpha,\beta} n_{\alpha\beta}!} p_{\alpha\beta}^{n_{\alpha\beta}} \right],$$

where

$$(3.9)\quad L_{a,b,c,d}(n):\ \begin{array}{l} N_{1.} = h-1-\delta_{1a}-\delta_{1b},\ \ N_{2.} = i-h-1-\delta_{2a}-\delta_{2b}, N_{3.} = n-i-\delta_{3a}-\delta_{3b}, \\ N_{.1} = h-1-\delta_{1c}-\delta_{1d},\ \ N_{.2} = k-j-1-\delta_{2c}-\delta_{2d}, N_{.3} = n-k-\delta_{3c}-\delta_{3d}. \end{array}$$

Thus we have

Theorem 3.2. The joint density function of the (h,j)th and the (i,k)th order statistics, i.e., the joint density function of x_h, x_i, y_j and y_k $(h < i,\ j < k)$ is the sum of (3.3), (3.7), and (3.8).

3.3. The joint density function of $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$.

In this extreme case, we have a very simple expression. Denote the set of all $n!$ permutations of $(1,2,\ldots,n)$ by $\mathcal{P}$. Then it is easy to see that the desired joint density function can be expressed in the following two alternative forms:

$$n! \sum_{(i_1,\ldots,i_n)\in\mathcal{P}} f(x_1,y_{i_1})f(x_2,y_{i_2})\cdots f(x_{j_n},y_n). \tag{3.10}$$

or

$$n! \sum_{(j_1,\ldots,j_n)\in\mathcal{P}} f(x_{j_1},y_1)f(x_{j_2},y_2)\cdots f(x_{j_n},y_n). \tag{3.11}$$

When x^* and y^* are independent, we have, of course,

$$n!\, f_1(x_1)\cdots f_1(x_n)\cdot n!\, f_2(y_1)\cdots f_2(y_n)\ . \tag{3.12}$$

3.4. Remark on the general case.

Let

$$1 \leqq n_1 \leqq n_2 \leqq \cdots \leqq n_{k_1} \leqq n \quad \text{and} \quad 1 \leqq m_1 \leqq m_2 \leqq \cdots \leqq m_{k_2} \leqq n$$

and consider the joint d.f. of the order statistics (x_{n_i}, y_{m_j}), $i=1,\ldots,k_1$; $j=1,\ldots,k_2$, i.e., the joint d.f. of $x_{n_1},\ldots,x_{n_{k_1}}$ and $y_{m_1},\ldots,y_{m_{k_2}}$, where $k_1 \geqq k_2$. Though we can, in principle, write the whole expression from a geometrical argument, it is terribly cumbersome, and so is omitted here. In this subsection, we make a short remark on the cases to be considered according to the number of sample points determining our order statistics. The simplest case is the one when our order statistics are determined by k_1 different sample points in such a way that k_2 sample points coincide with our order statistics and on the other hand, each of (k_1-k_2) sample points has only one component giving one of our order statistic in component. Another extreme case is as follows: If $k_1+k_2 \leqq n$, then this is the case when our order statistics are determined by k_1+k_2 different sample points, in which either component of each point coincides with one of the order statistics in component and the other component is different from anyone of our order statistics in component. If $k_1+k_2 \geqq n$, then the extreme case happens when our order statistics are determined by n sample points in such a way that (k_1+k_2-n) sample points are consistent with some (k_1+k_2-n) points of our order statistics and each of the remaining $2n-k_1-k_2$ sample points has only one component consistent with one of our order statistics in component. In the intermediate cases between these two extreme cases, our order statistics are determined by $k_1+1,\ldots,k_1+k_2-1$ (if $k_1+k_2 \leqq n$) or $n-1$ (if $k_1+k_2 \geqq n$), sample points. There are k_2+1 or $n-k_1+1$ cases in total. If we separately evaluate the contributions due to these cases to the whole joint d.f. and then sum, we obtain the desired exact joint d.f.

4. Bivariate range and its mean and variance.

Definition 3. Let x_n and x_1 be the maximum and the minimum of the x-components, y_n and y_1 be the maximum and the minimum of the y-components, in a sample of size n drawn from a bivariate population. Then the quantity

$$\begin{aligned} R(x,y) &= (x_n-x_1)(y_n-y_1) \\ &= R_1(x)R_2(y) \end{aligned} \tag{4.1}$$

is called a bivariate range in the sample. In other words, the bivariate range in the sample is defined as the product of ranges, $R_1(x)$ and $R_2(y)$, of component variables.

$R(x,y)$ represents the area of the smallest rectangle which includes all the sample points in the (x,y)-plane, and as in the univariate case, is regarded as a measure of dispersion in the sample.

4.1. Mean of $R(x,y)$.

Throughout this and next subsections, we assume the existence of integrals or of moments appeared in the calculation. First we derive the expression for the mean of $R(x,y)$ in the form convenient for the numerical computation which is the generalization of Tippett's formula in the univariate case [8], which was first treated by Mardia [2,3]. In Theorem 3.1, we have the joint d.f. of x_n, x_1, y_n and y_1. With the same notation as in Theorem 3.1, the expression to be evaluated is

$$E\{R(x,y)\} = E\{(x_n-x_1)(y_n-y_1)\}$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{x_n}\int_{-\infty}^{\infty}\int_{-\infty}^{y_n}(x_n-x_1)(y_n-y_1)\varphi_n(x_1,y_1,x_n,y_n) \cdot dy_1 dy_n dx_1 dx_n \tag{4.2}$$

$$\equiv A_1 + A_2 + A_3 ,$$

where

$$A_1 = n(n-1)(n-2)(n-3)\int_{-\infty}^{\infty}\int_{-\infty}^{x_n}\int_{-\infty}^{\infty}\int_{-\infty}^{y_n}(x_n-x_1)(y_n-y_1)P^{n-4} \cdot p_1 q_1 p_n q_n dy_1 dy_n dx_1 dx_n , \tag{4.3}$$

$$A_2 = n(n-1)(n-2)\int_{-\infty}^{\infty}\int_{-\infty}^{x_n}\int_{-\infty}^{\infty}\int_{-\infty}^{y_n}(x_n-x_1)(y_n-y_1)P^{n-3} \cdot \{p_n q_n f(1,1)+p_1 q_n f(1,n)+p_n q_1 f(n,1)+p_1 q_1 f(n,n)\} \cdot dy_1 dy_n dx_1 dx_n , \tag{4.4}$$

$$A_3 = n(n-1)\int_{-\infty}^{\infty}\int_{-\infty}^{x_n}\int_{-\infty}^{\infty}\int_{-\infty}^{y_n}(x_n-x_1)(y_n-y_1)P^{n-2}\{f(1,1)f(n,n)+ f(1,n)f(n,1)\}dy_1 dy_n dx_1 dx_n . \tag{4.5}$$

In the computation below, we use the following lemmas:

Lemma 4.1. *Suppose that* $V(y_n)$ *has continuous derivatives and vanishes at* $y_n = -\infty$. *Then, for any positive integer* m,

$$\int_{-\infty}^{\infty}\{F(x_n,y_n)-F(x_1,y_n)\}^m p_n\, V(y_n)dy_n$$
$$= \frac{1}{m+1}\int_{-\infty}^{\infty}[\{F_1(x_n)-F_1(x_1)\}^{m+1}-\{F(x_n,y_n)-F(x_1,y_n)\}^{m+1}]\,\frac{dV(y_n)}{dy_n}\,dy_n. \tag{4.6}$$

Proof. Put

$$W = \int_{F(n,n)-F(1,n)}^{F_1(n)-F_1(1)} g^m dg = \frac{1}{m+1}[\{F_1(n)-F_1(1)\}^{m+1}-\{F(n,n)-F(1,n)\}^{m+1}],$$

where we used the abbreviated notations: $F_1(n) \equiv F_1(x_n)$, $F_1(1) \equiv F_1(x_1)$, $F(n,n) \equiv F(x_n,y_n)$, $F(1,n) \equiv F(x_1,y_n)$. Then

$$\int_{-\infty}^{\infty}\{F(n,n)-F(1,n)\}^m p_n V(y_n)dy_n = -\int_{-\infty}^{\infty}\frac{dW}{dy_n}V(y_n)dy_n = -[WV(y_n)]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} W\frac{dV(y_n)}{dy_n}dy_n$$

$$= \int_{-\infty}^{\infty} W\frac{dV(y_n)}{dy_n}dy_n ,$$

which is (4.6).

Lemma 4.2. *Let* $V(y_n)$ *be a function of* y_n *defined in Lemma 4.1. Then*

$$\int_{-\infty}^{\infty} f(x,y_n)V(y_n)dy_n = \int_{-\infty}^{\infty}\{\int_{y_n}^{\infty} f(x,y)dy\}\frac{dV(y_n)}{dy_n}dy_n \tag{4.7}$$

The proof is similar as in Lemma 4.1.

If we change the role of y_n to that of x_n, then we have similar lemmas. Now let us start with the evaluation of A_1 in (4.3.).

$$A_{11} = n(n-1)(n-2)(n-3)\int_{-\infty}^{\infty}\int_{-\infty}^{y_n}(y_n-y_1)P^{n-4}p_1p_nq_1q_ndy_1dy_n$$

$$= n(n-1)(n-2)(n-3)\sum_{j=0}^{n-4}(-1)^j\binom{n-4}{j}\int_{-\infty}^{\infty}\int_{-\infty}^{y_n}(y_n-y_1)\{F(n,1)-F(1,1)\}^j$$

$$\cdot\{F(n,n)-F(1,n)\}^{n-4-j}p_1p_nq_1q_ndy_1dy_n .$$

If we put

$$V(y_n) = \int_{-\infty}^{y_n} (y_n - y_1)\{F(n,1) - F(1,1)\}^j p_1 q_1 q_n \, dy_1 \, ,$$

then, by Lemma 4.1,

$$A_{11} = n(n-1)(n-2)(n-3) \sum_{j=0}^{n-4} (-1)^j \binom{n-4}{j} \frac{1}{n-3-j} \int_{-\infty}^{\infty} [\{F_1(n) - F_1(1)\}^{n-3-j} - \{F(n,n) - F(1,n)\}^{n-3-j}] \frac{dV(y_n)}{dy_n} \, dy_n .$$

Since

$$\frac{dV(y_n)}{dy_n} = \int_{-\infty}^{y_n} \{F(n,1) - F(1,1)\}^j p_1 q_1 q_n \, dy_1$$

$$+ \int_{-\infty}^{y_n} (y_n - y_1)\{F(n,1) - F(1,1)\}^j \{q_1 f(n,n) + q_n f(1,n)\} p_1 \, dy_1$$

$$= \frac{1}{j+1} \int_{-\infty}^{y_n} \{F(n,1) - F(1,1)\}^{j+1} \{q_1 f(n,n) + q_n f(1,n) + q_1 f(n,1) + q_n f(1,1)\} dy_1$$

$$+ \frac{1}{j+1} \int_{-\infty}^{y_n} (y_n - y_1)\{F(n,1) - F(1,1)\}^{j+1} \{f(1,1) f(n,n) + f(1,n) f(n,1)\} dy_1 ,$$

A_{11} can be reduced to

$$A_{11} = n(n-1) \int_{-\infty}^{\infty} \int_{-\infty}^{y_n} [\{F_1(n) - F_1(1)\}^{n-2} - \{F(n,n) - F(1,n)\}^{n-2} - \{F_1(n) - F_1(1) - F(n,1) + F(1,1)\}^{n-2} + P^{n-2}] \times$$

$$\times [\{q_1 f(n,n) + q_n f(1,n) + q_1 f(n,1) + q_n f(1,1)\} + (y_n - y_1)\{f(1,1) f(n,n) + f(1,n) f(n,1)\}] dy_1 dy_n . \quad (4.8)$$

For A_2, let us evaluate

$$A_{21} = n(n-1)(n-2)\int_{-\infty}^{\infty}\int_{-\infty}^{y_n}(y_n-y_1)P^{n-3}\{q_n f(1,1)+q_1 f(n,1)\}p_n\,dy_1\,dy_n$$

$$+\, n(n-1)(n-2)\int_{-\infty}^{\infty}\int_{-\infty}^{y_n}(y_n-y_1)P^{n-3}\{q_n f(1,n)+q_1 f(n,n)\}r_1\,dy_1\,dy_n$$

$$= A_{21,1} + A_{21,2}.$$

Using Lemma 4.1 similarly as for A_{11}, we have

$$A_{21,1} = n(n-1)\int_{-\infty}^{\infty}\int_{-\infty}^{y_n}[\{F_1(n)-F_1(1)-F(n,1)+F(,1,)\}^{n-2}-P^{n-2}] \times$$

$$(4.9) \qquad \times\, [\{q_n f(1,1)+q_1 f(n,1)\}+(y_n-y_1)\{f(1,1)f(n,n)+f(1,n)f(n,1)\}]$$

$$\cdot\, dy_1\,dy_n\ .$$

By integration by parts with respect to y_1, $A_{21,2}$ can be reduced to

$$A_{21,2} = n(n-1)\int_{-\infty}^{\infty}\int_{-\infty}^{y_n}[\{F(n,n)-F(1,n)\}^{n-2}-P^{n-2}] \times$$

$$(4.10) \qquad \times\, [\{q_n f(1,n)+q_1 f(n,n)\}+(y_n-y_1)\{f(1,1)f(n,n)+f(1,n)f(n,1)\}]dy_1\,dy_n.$$

Now we insert the results (4.8), (4.9), and (4.10) into (4.2) and then we have

$$E\{R(x,y)\} = \int_{-\infty}^{\infty}\int_{-\infty}^{x_n}(x_n-x_1)(A_{11}+A_{21,1}+A_{21,2})dx_1\,dx_n + A_3$$

$$= n(n-1)\int_{-\infty}^{\infty}\int_{-\infty}^{x_n}(x_n-x_1)dx_1\,dx_n \times$$

$$(4.11) \qquad \times \Big[\{F_1(n)-F_1(1)\}^{n-2}\int_{-\infty}^{\infty}\int_{-\infty}^{y_n}(y_n-y_1)\{f(1,1)f(n,n)+ f(1,n)f(n,1)\}dy_1\,dy_n$$

$$+\int_{-\infty}^{\infty}[\{F_1(n)-F_1(1)\}^{n-2}-\{F(n,n)-F(1,n)\}^{n-2}](\int_{-\infty}^{y_n}\{q_nf(1,1)+$$

$$+q_1f(n,1)\}dy_1)dy_n$$

$$+\int_{-\infty}^{\infty}[\{F_1(n)-F_1(1)\}^{n-2}-\{F_1(n)-F_1(1)-F(n,1)+F(1,1)\}^{n-2}]$$

$$\cdot(\int_{y_1}^{\infty}\{q_nf(1,n)+q_1f(n,n)\}dy_n)dy_1\Big].$$

Since

$$\int_{-\infty}^{\infty}\int_{-\infty}^{y_n}(y_n-y_1)\{f(1,1)f(n,n)+f(1,n)f(n,1)\}dy_1dy_n$$

$$=\int_{-\infty}^{\infty}\{(\int_{y_n}^{\infty}f(x_n,y)dy)(\int_{-\infty}^{y_n}f(x_1,y)dy)$$

$$+(\int_{y_n}^{\infty}f(x_1,y)dy)(\int_{-\infty}^{y_n}f(x_n,y)dy)\}dy_n$$

by Lemma 4.2 and

$$\int_{-\infty}^{y_n}\{q_nf(1,1)+q_1f(n,1)\}dy_1=(\int_{-\infty}^{y_n}f(x_n,y)dy)(\int_{-\infty}^{y_n}f(x_1,y)dy)\ ,$$

$$\int_{y_1}^{\infty}\{q_nf(1,n)+q_1f(n,n)\}dy_n=(\int_{y_1}^{\infty}f(x_n,y)dy)(\int_{y_1}^{\infty}f(x_1,y)dy)\ ,$$

and since y_n and y_1 are dummy variables, we can write $E\{R(x,y)\}$ in the form,

$$
\begin{aligned}
E\{R(x,y)\} = {} & n(n-1)\int_{-\infty}^{\infty}\int_{-\infty}^{x_n}(x_n-x_1)dx_1dx_n \times \\
& \times\Big[\{F_1(n)-F_1(1)\}^{n-2}\int_{-\infty}^{\infty}\{(\int_z^{\infty} f(x_n,y)dy)(\int_{-\infty}^{z} f(x_1,y)dy) + \\
& \qquad (\int_z^{\infty} f(x_1,y)dy)(\int_{-\infty}^{z} f(x_n,y)dy)\}dz \\
& + \int_{-\infty}^{\infty}[\{F_1(n)-F_1(1)\}^{n-2}-\{F(n,z)-F(1,z)\}^{n-2}] \, . \\
& \qquad \cdot(\int_{-\infty}^{z}(x_n,y)dy)(\int_{-\infty}^{z} f(x_1,y)dy)dz \\
& + \int_{-\infty}^{\infty}[\{F_1(n)-F_1(1)\}^{n-2}-\{F_1(n)-F_1(1)-F(n,z)+F(1,z)\}^{n-2}] \\
& \qquad \cdot(\int_z^{\infty} f(x_n,y)dy)(\int_z^{\infty} f(x_1,y)dy)dz\Big].
\end{aligned}
\tag{4.12}
$$

Hence, after the interchange of integration-order, we have

$$
\begin{aligned}
E\{R(x,y)\} = {} & \int_{-\infty}^{\infty}dz\Big[n(n-1)\int_{-\infty}^{\infty}\int_{-\infty}^{x_n}(x_n-x_1)\{F_1(x_n)-F_1(x_1)\}^{n-2} \\
& \qquad \cdot f_1(x_n)f_1(x_1)dx_1dx_n \\
& -n(n-1)\int_{-\infty}^{\infty}\int_{-\infty}^{x_n}(x_n-x_1)\{F(x_n,z)-F(x_1,z)\}^{n-2} \\
& \qquad \cdot(\int_{-\infty}^{z} f(x_n,y)dy)(\int_{-\infty}^{z} f(x_1,y)dy)dx_1dx_n \\
& -n(n-1)\int_{-\infty}^{\infty}\int_{-\infty}^{x_n}(x_n-x_1)\{F_1(x_n)-F_1(x_1)-F(x_n,z)+F(x_1,z)\}^{n-2} \\
& \qquad \cdot(\int_z^{\infty} f(x_n,y)dy)(\int_z^{\infty} f(x_1,y)dy)dx_1dx_n\Big] \\
= {} & \int_{-\infty}^{\infty}(I_1 - I_2 - I_3)dz \, .
\end{aligned}
\tag{4.13}
$$

In this stage, we can apply Tippett's derivation [8] to I_1, I_2, and I_3, that is,

$$I_1 = \int_{-\infty}^{\infty} [1-\{1-F_1(u)\}^n-\{F_1(u)\}^n]du \;, \tag{4.14}$$

$$I_2 = \int_{-\infty}^{\infty} [\{F_2(z)\}^n-F_2(z)-F(u,z)\}^n-\{F(u,z)\}^n]du \;, \tag{4.15}$$

$$I_3 = \int_{-\infty}^{\infty} [\{1-F_2(z)\}^n-\{1-F_2(z)-F_1(u)+F(u,z)\}^n-\{F_1(u)-F(u,z)\}^n]du. \tag{4.16}$$

Thus we finally have

$$\begin{aligned} E\{R(x,y)\} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [1-\{1-F_1(u)\}^n-\{1-F_2(z)\}^n-\{F_1(u)\}^n-\{F_2(z)\}^n+\{F(u,z)\}^n \\ &\quad + \{F_1(u)-F(u,z)\}^n+\{F_2(z)-F(u,z)\}^n+\{1-F_1(u)-F_2(z)+F(u,z)\}^n]dudz \;. \end{aligned} \tag{4.17}$$

If x and y are independent,

$$\begin{aligned} E\{R(x,y)\} &= \int_{-\infty}^{\infty} [1-\{1-F_1(u)\}^n-\{F_1(u)\}^n]du \cdot \int_{-\infty}^{\infty} [1-\{1-F_2(z)\}^n-F_{\{}(z)\}^n]dz \\ &= E\{R(x)\}E\{R(y)\} \;. \end{aligned}$$

which is the product of Tippett's formulas for each components.

4.2. Variance of $R(x,y)$.

It is sufficient to evaluate $E\{R^2(x,y)\}$, i.e.,

$$\begin{aligned} E\{R^2(x,y)\} &= E\{(x_n-x_1)^2(y_n-y_1)^2\} \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{x_n}\int_{-\infty}^{\infty}\int_{-\infty}^{y_n}(x_n-x_1)^2(y_n-y_1)^2\varphi_n(x_1,y_1,x_n,y_n)dy_1dy_ndx_1dx_n \\ &\equiv B_1+ B_2+ B_3 \;. \end{aligned} \tag{4.18}$$

In the same manner as in the case of the mean, we have

$$\begin{aligned} B_{11} &= n(n-1)(n-2)(n-3)\int_{-\infty}^{\infty}\int_{-\infty}^{y_n}(y_n-y_1)^2P^{n-4}p_1p_nq_1q_n\,dy_1dy_n \\ &= n(n-1)\int_{-\infty}^{\infty}\int_{-\infty}^{y_n}[\{F_1(n)-F_1(1)\}^{n-2}-\{F(n,n)-F(1,n)\}^{n-2}-\{F_1(n)-F_1(1)- \\ &\qquad F(n,1)+F(1,1)\}^{n-2}+P^{n-2}]\times \\ &\quad\times[2q_1q_n+2(y_n-y_1)\{q_1f(n,n)+q_1f(n,1)+q_nf(1,n)+q_nf(1,1)\} \\ &\qquad+(y_n-y_1)^2\{f(1,1)f(n,n)+f(1,n)f(n,1)\}]dy_1dy_n. \end{aligned}\tag{4.19}$$

$$\begin{aligned} B_{21} &= n(n-1)(n-2)\int_{-\infty}^{\infty}\int_{-\infty}^{y_n}(y_n-y_1)^2P^{n-3}[\{q_nf(1,1)+q_1f(n,1)\}p_n+ \\ &\qquad\{q_nf(1,n)+q_1f(n,n)\}p_1]dy_1dy_n \\ &= n(n-1)\int_{-\infty}^{\infty}\int_{-\infty}^{y_n}\{F_1(n)-F_1(1)-F(n,1)+F(1,1)\}^{n-2}\times \\ &\quad\times[2(y_n-y_1)\{q_nf(1,1)+q_1f(n,1)\}+(y_n-y_1)^2\{f(1,1)f(n,n)+f(1,n) \\ &\qquad f(n,1)\}]dy_1dy_n \\ &\quad+n(n-1)\int_{-\infty}^{\infty}\int_{-\infty}^{y_n}\{F(n,n)-F(1,n)\}^{n-2}[2(y_n-y_1)\{q_nf(1,n)+q_1f(n,n)\}+ \\ &\qquad+(y_n-y_1)^2\{f(1,1)f(n,n)+f(1,n)f(n,1)\}]dy_1dy_n \\ &\quad-n(n-1)\int_{-\infty}^{\infty}\int_{-\infty}^{y_n}P^{n-2}[2(y_n-y_1)\{q_nf(1,1)+q_nf(1,n)+q_1f(n,1)+q_1f(n,n)\}+ \\ &\qquad+(y_n-y_1)^2\{f(1,1)f(n,n)+f(1,n)f(n,1)\}]dy_1dy_n \end{aligned}\tag{4.20}$$

Inserting (4.19) and (4.20) into (4.18), we have

$$\begin{aligned} E\{R^2(x,y)\} &= \int_{-\infty}^{\infty}\int_{-\infty}^{x_n}(x_n-x_1)^2(B_{11}+B_{21})dx_1dx_n+B_3=\int_{-\infty}^{\infty}\int_{-\infty}^{x_n}(x_n-x_1)^2I(x_1,x_n) \\ &\qquad dx_1dx_n, \end{aligned}\tag{4.21}$$

where

$$
\begin{aligned}
I(x_1,x_n) = & \int_{-\infty}^{\infty}\int_{-\infty}^{y_n}(y_n-y_1)^2\{F_1(n)-F_1(1)\}^{n-2}\{f(1,1)f(n,n)+f(1,n)f(n,1)\} \\
& \qquad\qquad \cdot dy_1dy_n \\
& + 2\int_{-\infty}^{\infty}\int_{-\infty}^{y_n}(y_n-y_1)[\{F_1(n)-F_1(1)\}^{n-2}-\{F(n,n)-F(1,n)\}^{n-2}] \\
& \qquad\qquad \cdot\{q_nf(1,1)+q_1f(n,1)\}dy_1dy_n \\
& + 2\int_{-\infty}^{\infty}\int_{-\infty}^{y_n}(y_n-y_1)[\{F_1(n)-F_1(1)\}^{n-2}-\{F_1(n)-F_1(1)-F(n,1)+F(1,1)\}^{n-2}] \\
& \qquad\qquad \cdot\{q_nf(1,n)+q_1f(n,n)\}dy_1dy_n \\
& + 2\int_{-\infty}^{\infty}\int_{-\infty}^{y_n}[\{F_1(n)-F_1(1)\}^{n-2}-\{F(n,n)-F(1,n)\}^{n-2}-\{F_1(n)- \\
& \qquad\qquad -F_1(1)-F(n,1)+F(1,1)\}^{n-2}+P^{n-2}]q_1q_ndy_1dy_n.
\end{aligned}
\tag{4.22}
$$

Let us introduce the following notation,

$$a_i \equiv \int_{y_n}^{\infty} f(x_i,y)dy,\ i=1,n\ ;\ b_i = \int_{-\infty}^{y_1} f(x_i,y)dy,\ i=1,n\ ,$$

and let us continue to eliminate (y_n-y_1) and $(y_n-y_1)^2$. It is easy to see that by Lemma 4.2 and by integration by parts, we obtain

$$
\begin{aligned}
\int_{-\infty}^{\infty}\int_{-\infty}^{y_n}(y_n-y_1)^2\{f(1,1)f(n,n)+f(1,n)f(n,1)\}dy_1dy_n = & \\
= 2\int_{-\infty}^{\infty}\int_{-\infty}^{y_n}(a_nb_1+a_1b_n)dy_1dy_n\ , &
\end{aligned}
\tag{4.23}
$$

$$\int_{-\infty}^{y_n}(y_n-y_1)\{q_nf(1,1)+q_1f(n,1)\}dy_1 = \int_{-\infty}^{y_n}(q_nb_1+q_1b_n+b_1b_n)dy_1\ , \tag{4.24}$$

$$\int_{y_1}^{\infty}(y_n-y_1)\{q_nf(1,n)+q_1f(n,n)\}dy_n = \int_{y_1}^{\infty}\{q_na_1+q_1a_n+a_1a_n]dy_n \tag{4.25}$$

Inserting these results into (4.22), we can write $I(x_1,x_n)$ in the form,

$$
\begin{aligned}
I(x_1,x_n) = 2\int_{-\infty}^{\infty}\int_{-\infty}^{y_n} & [(F_1(n)-F_1(1))^{n-2}(q_nq_1+q_na_1+q_1a_n+q_nb_1+q_1b_n+ \\
& \qquad a_1a_n+a_1b_n+a_nb_1+b_1b_n) \\
& -(F(n,n)-F(1,n))^{n-2}(q_nq_1+q_nb_1+q_1b_n+b_1b_n) \\
& -(F_1(n)-F_1(1)-F(n,1)+F(1,1))^{n-2}(q_nq_1+q_na_1+q_1a_n+a_1a_n) \\
& + P^{n-2}q_nq_1]dy_1dy_n .
\end{aligned}
\tag{4.26}
$$

If we interchange the order of integration between y_1, y_n and x_1, x_n, we have

$$
\begin{aligned}
E(R^2(x,y)) = 2\int_{-\infty}^{\infty}\int_{-\infty}^{y_n} dy_1dy_n\cdot n(n-1) & \int_{-\infty}^{\infty}\int_{-\infty}^{x_n}(x_n-x_1)^2dx_1dx_n \times \\
\times [(F_1(n)-F_1(1))^{n-2} & f_1(x_n)f_1(x_1)-(F(n,n)-F(1,n))^{n-2} \cdot \\
& \cdot\left(\int_{-\infty}^{y_n}f(x_1,y)dy\right)\left(\int_{-\infty}^{y_n}f(x_n,y)dy\right) \\
-(F_1(n)-F_1(1) & -F(n,1)+F(1,1))^{n-2}\cdot \\
& \cdot\left(\int_{y_1}^{\infty}f(x_1,y)dy\right)\left(\int_{y_1}^{\infty}f(x_n,y)dy\right)+P^{n-2}q_nq_1] , \\
= 2\int_{-\infty}^{\infty}\int_{-\infty}^{y_n} & dy_1dy_n\cdot(J_1-J_2-J_3+J_4) .
\end{aligned}
\tag{4.27}
$$

Now we can use the Tippett formula for the variance in the univariate case to J_i, $i=1,2,3,4$.

$$J_1 = 2\int_{-\infty}^{\infty}\int_{-\infty}^{x_n}[1-\{F_1(n)\}^n-\{1-F_1(1)\}^n+\{F_1(n)-F_1(1)\}^n]dx_1dx_n, \tag{4.28}$$

$$J_2 = 2\int_{-\infty}^{\infty}\int_{-\infty}^{x_n}[\{F_2(n)\}^n-F(n,n)\}^n-\{F_2(n)-F(1,n)\}^n+\{F(n,n)-F(1,n)\}^n] \cdot dx_1dx_n, \tag{4.29}$$

$$J_3 = 2\int_{-\infty}^{\infty}\int_{-\infty}^{x_n}[\{1-F_2(1)\}^n-\{F_1(1)-F(1,1)\}^n-\{1-F_1(1)-F_2(1)+F(1,1)\}^n +\{F_1(n)-F_1(1)-F(n,1)+F(1,1)\}^n]dx_1dx_n, \tag{4.30}$$

$$J_4 = 2\int_{-\infty}^{\infty}\int_{-\infty}^{x_n}[\{F_2(n)-F_2(1)\}^n-\{F(n,n)-F(n,1)\}^n-\{F_2(n)-F_2(1)-F(1,n)+F(1,1)\}^n +\{F(n,n)-F(n,1)-F(1,n)+F(1,1)\}^n] \cdot dx_1dx_n . \tag{4.31}$$

The final formula for the variance is obtained by inserting these results into (4.27), which is summarized, with the formula for the mean, in the

Theorem 4.1. *The mean and the variance of the bivariate range* $R(x,y) = (x_n-x_1)(y_n-y_1)$ *in a sample of size* n *drawn from a bivariate continuous population with d.f.* $F(x,y)$ *can be evaluated by the following formulas*:

$$E\{R(x,y)\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[(1-\{F_1(x)\}^n-\{1-F_1(x)\}^n)-(\{F_2(y)\}^n-\{F(x,y)\}^n -\{F_2(y)-F(x,y)\}^n)-(\{1-F_2(y)\}^n-\{F_1(x)-F(x,y)\}^n-\{1-F_1(x)- F_2(y)+F(x,y)\}^n)]dxdy , \tag{4.32}$$

and

$$\mathrm{Var}\{R(x,y)\} = E\{R^2(x,y)\}-E^2\{R(x,y)\},$$

$$\begin{aligned}
E\{R^2(x,y)\} &= 4\int_{-\infty}^{\infty}\int_{-\infty}^{v} du dv \cdot \int_{-\infty}^{\infty}\int_{-\infty}^{t} ds dt \times \\
&\times [(1-\{F_1(v)\}^n-\{1-F_1(u)\}^n + \{F_1(v)-F_1(u)\}^n) \\
&\quad -(\{F_2(t)\}^n-\{F(v,t)\}^n-\{F_2(t)-F(u,t)\}^n+\{F(v,t)-F(u,t)\}^n) \\
&\quad -(\{1-F_2(s)\}^n-\{F_1(v)-F(v,s)\}^n-\{1-F_2(s)-F_1(u)+F(u,s)\}^n+ \\
&\qquad \{F_1(v)-F_1(u)-F(v,s)+F(u,s)\}^n) \\
&\quad + (\{F_2(t)-F_2(s)\}^n-\{F(v,t)-F(v,s)\}^n-\{F_2(t)-F_2(s)-F(u,t)+F(u,s)\}^n \\
&\qquad +\{F(v,t)-F(v,s)-F(u,t)+F(u,s)\}^n)],
\end{aligned} \tag{4.33}$$

where F_1 *and* F_2 *are the marginal d.f.'s of* x *and* y, *respectively*.

5. Asymptotic distributions of the quantiles in the two-dimensional case.

Suppose that we have a sample size n taken from a bivariate distribution of continuous type with d.f. $F(x,y)$, and the d.f. $f(x,y)$. As before, let us denote the marginal c.d.f.'s and d.f.'s of x and y by $F_1(x)$, $f_1(x)$ and $F_2(y)$, $f_2(y)$, respectively.

Let $\xi_p \equiv \xi$ be the quantile of order p of the marginal d.f. of x, i.e., the root of the equation: $F_1(\xi_p) = p$, $(0 < o < 1)$ and similarly we shall denote the quantile of order q of the marginal distribution of y by $\eta_q \equiv \eta$, i.e., the root of the equation: $F_2(\eta_q) = q$.

Definition 4. The quantity, (ξ_p,η_q) is called *the combined quantile of order* (p,q) *in the bivariate distribution*.

Now we arrange n values of each component in the sample in acending order of magnitude separately.

$$x_1 \leqq x_2 \leqq \cdots \leqq x_n \; ; \quad y_1 \leqq y_2 \leqq \cdots \leqq y_n \; .$$

In order to avoid trivial complications, we assume in the sequel that both np and nq are not integers. Then

Definition 5. The sample combined quantile, $(z^{(1)}, z^{(2)})$, corresponding to (ξ_p,η_q) is uniquely defined as *the bivariate order statistic*, $(x_{\alpha+1}, y_{\beta+1})$, where α and β are the greatest integers $\leqq np$ and $\leqq nq$, i.e., $\alpha = [np]$ and $\beta = [nq]$, (Gauss' symbol), respectively.

In a special case when $p=q=\frac{1}{2}$, i.e., in the case of, so called, the bivariate median, Mood [4] gave its asymptotic distribution. In this section, we treat the general case of p and q. To do this, we need the assumption that *in some neighborhood of the point* $(x,y) = (\xi_p,\eta_q)$, $f(x,y)$ *is continuous and has continuous derivatives as a function of two variables, and* $f_1(\xi_p,) \neq 0$ *and* $f_2(\eta_q) \neq 0$.

We have already obtained the exact d.f. of $(z^{(1)}, z^{(2)})$ in Theorem 2.1, i.e., $\phi_n(x_{\alpha+1}, y_{\beta+1})$ in (2.14). In this expression, the sum (2.8) is due to the case when $(x_{\alpha+1}, y_{\beta+1})$ is a sample point, and other four sums, (2.10), (2.11), (2.12) and (2.13) are due to the case when the order statistics is determined by two different sample points. For the investigation of the asymptotic behavior of $(z^{(1)}, z^{(2)})$, it is to be noted that (2.8) is of smaller order than others with respect to n. In fact, we have in (2.8) with $i=\alpha+1$ and $j=\beta+1$,

$$C(1,n_{11},\alpha-n_{11},\beta-n_{11};n) = n\binom{n}{\alpha}\binom{\alpha}{n_{11}}\binom{n-\alpha}{\beta-n_{11}}\left\{\frac{n_{11}-nP'_{11}}{n}+(1-p-q+P'_{11}) + O\left(\frac{1}{n}\right)\right\},$$

where $P'_{11} = F(\xi,\eta)$. On the other hand, for example, in (2.10) with $i=\alpha+1$ and $j=\beta+1$,

$$C(1,1,n_{11},\alpha-1-n_{11},\beta-1-n_{11};n)=n^2\binom{n}{\alpha}\binom{\alpha}{n_{11}}\binom{n-\alpha}{\beta-n_{11}}\left\{\frac{n_{11}-nP'_{11}}{n}-(p-P'_{11})+O\left(\frac{1}{n}\right)\right\} \cdot\left\{\frac{n_{11}-nP'_{11}}{n}-(q-P'_{11})+O\left(\frac{1}{n}\right)\right\}.$$

Since we have only to consider the values of n_{11} such that $|n_{11}-nP'_{11}|n^{-1}\to 0$ according to the law of large numbers, the term (2.8) is of order $\frac{1}{n}$ in comparison with the (2.10). (2.10), (2.11), (2.12), and (2.13) are of the same order. Accordingly, (2.8) may be neglected in the asymptotic form of the d.f. of $(z^{(1)}, z^{(2)})$.

Now consider the sum (2.10) with $i=\alpha+1$ and $j=\beta+1$, i.e.,

$$(5.1)\quad g_n^*(Z^{(1)},Z^{(2)}) = F(x_{\alpha+1}\|y_{\beta+1})F(y_{\beta+1}\|x_{\alpha+1})\cdot\sum_{n_{11}=0}^{m-1} C(1,1,n_{11},\alpha-1-n_{11},\beta-1-n_{11})P_{11}^{n_{11}}P_{12}^{\alpha-1-n_{11}}P_{21}^{\beta-1-n_{11}}P_{22}^{n+n_{11}-\alpha-\beta},$$

where $m = \min(\alpha,\beta)$. Noticing that, in the marginal case,

$$(5.2) \qquad u = \sqrt{\frac{n}{p(1-p)}}\, f_1(\xi)(Z^{(1)}-\xi) \quad \text{and} \quad v = \sqrt{\frac{n}{q(1-q)}}\, f_2(\eta)(Z^{(2)}-\eta)$$

are respectively distributed in the limit according to the standard normal distribution [1;p.369], we shall consider the combined random variable (u,v). Making the transformation (5.2), the expression to be evaluated is

$$(5.3) \qquad \begin{aligned} q_n'(u,v) &= \frac{1}{f_1(\xi)f_2(\eta)} \frac{\sqrt{p(1-p)q(1-q)}}{n}\, g^*\Big(\xi + \sqrt{\frac{p(1-p)}{n}}\frac{1}{f_1(\xi)}\, u, \\ &\qquad\qquad \eta + \sqrt{\frac{q(1-q)}{n}}\frac{1}{f_2(\eta)}\, v\Big) \\ &= \sqrt{p(1-p)q(1-q)}\;\; G_{on} \sum_{n_{11}=0}^{m-1} G_{1n}G_{2n}, \end{aligned}$$

where

$$(5.4) \qquad G_{on} = \frac{F(x_{\alpha+1}\|y_{\beta+1})F(y_{\beta+1}\|x_{\alpha+1})}{f_1(\xi)f_2(\eta)},$$

$$(5.5) \qquad G_{1n} = \frac{(n-1)!}{n_{11}!(\alpha-1-n_{11})!(\beta-1-n_{11})!(n+n_{11}-\alpha-\beta)!} A_{11}^{n_{11}} A_{12}^{\alpha-1-n_{11}} A_{21}^{\beta-1-n_{11}} A_{22}^{n+n_{11}-\alpha-\beta},$$

$$(5.6) \qquad \begin{aligned} G_{2n} &= \left[\frac{F(s,t)}{A_{11}}\right]^{n_{11}} \left[\frac{F_1(s)-F(s,t)}{A_{12}}\right]^{\alpha-1-n_{11}} \left[\frac{F_2(t)-F(s,t)}{A_{21}}\right]^{\beta-1-n_{11}} \\ &\quad \cdot \left[\frac{1-F_1(s)-F_2(t)+F(s,t)}{A_{22}}\right]^{n+n_{11}-\alpha-\beta}, \end{aligned}$$

where

$$A_{11} \equiv F(\xi,\eta) \ , \quad A_{12} \equiv F_1(\xi)-F(\xi,\eta) = p-F(\xi,\eta),$$

$$A_{21} \equiv F_2(\eta)-F(\xi,\eta)=q-F(\xi,\eta), \ A_{22} \equiv 1-F_1(\xi)-F_2(\eta)+F(\xi,\eta)=1-p-q+F(\xi,\eta)$$

and

$$s = \xi + \sqrt{\frac{p(1-p)}{n}}\,\frac{1}{f_1(\xi)}\,u \quad \text{and} \quad t=\eta+\sqrt{\frac{q(1-q)}{n}}\,\frac{1}{f_2(\eta)}\,v \ .$$

In order to investigate the behavior of the distribution of the (u,v) for large n, we shall evaluate G_{on}, G_{1n} and G_{2n} separately for any fixed u and v. First of all,

$$G_{on} = hh^*[1 + O(1/\sqrt{n})] \ , \tag{5.7}$$

where

$$h = \int_{-\infty}^{\eta} f(\xi,y)dy/f_1(\xi) \ , \quad h^* = \int_{-\infty}^{\xi} f(x,\eta)dx/f_2(\eta) \ . \tag{5.8}$$

For G_{1n}, we use the normal approximation for the multinomial distribution. Using the notations

$$n_1= n_{11}, \ n_2 = \alpha-n_{11}-1, \ n_3= \beta-n_{11}-1, \ n_4= n+n_{11}-\alpha-\beta$$

and defining the variables

$$w_1= (n_1-nA_{11})/\sqrt{n} \ , \quad w_2= (n_2-nA_{12})/\sqrt{n} \ , \ w_3= (n_3-nA_{21})/\sqrt{n} \ , \tag{5.9}$$

we can write

$$G_{1n}= (n-1)[|M|/(2\pi)^3]^{\frac{1}{2}}\exp[-\tfrac{1}{2}\sum_{i,j=1}^{3} m_{ij}w_iw_j]\Delta w_1\Delta w_2\Delta w_3[1+O(1/\sqrt{n})], \tag{5.10}$$

where

$$m_{11}=\frac{1}{A_{11}}+\frac{1}{A_{22}}\ ,\ m_{22}=\frac{1}{A_{12}}+\frac{1}{A_{22}}\ ,\ m_{33}=\frac{1}{A_{21}}+\frac{1}{A_{22}}\ ,\ m_{ij}=m_{ji}=\frac{1}{A_{22}},\qquad (i\neq j),$$

and $|M|$ is the determinant of (m_{ij}), which is $1/(A_{11}A_{12}A_{21}A_{22})$ and $\Delta w_i=\frac{1}{\sqrt{n}}$, $i=1,2,3$. However, from the definition of α and β, we have

$$w_2=(\alpha-n_{11}-1-nA_{12})/\sqrt{n}=([np]-n_{11}-1-np+nA_{11})/\sqrt{n}=-w_1+O(1/\sqrt{n}). \tag{5.11}$$

$$w_3=-w_1+O(1/\sqrt{n})\ . \tag{5.12}$$

Therefore, we have the approximation to G_{1n} in the form

$$G_{1n}=[A_{11}A_{12}A_{21}A_{22}(2\pi)^3]^{-\frac{1}{2}}\exp\{-\tfrac{1}{2}k^2w_1^2\}\,\Delta w_1\{1+O(1/\sqrt{n})\}\ , \tag{5.13}$$

where

$$k^2=\frac{1}{A_{11}}+\frac{1}{A_{12}}+\frac{1}{A_{21}}+\frac{1}{A_{22}}\ . \tag{5.14}$$

Finally let us evaluate G_{2n} for large n. From our assumption on $f(x,y)$, we have the expansions:

$$\begin{aligned}F(s,t)&=A_{11}+\sqrt{\frac{p(1-p)}{n}}\,hu+\sqrt{\frac{q(1-q)}{n}}\,h^*v\\&\quad+\frac{1}{2n}\{p(1-p)c_{11}u^2+2\sqrt{p(1-p)q(1-q)}\,c_{12}uv+q(1-q)c_{22}v^2\}+o(1/n),\end{aligned} \tag{5.15}$$

$$F_1(s)=p+\sqrt{\frac{p(1-p)}{n}}\,u+\frac{1}{2}\,\frac{p(1-p)}{n}\,b_1u^2+o(1/n)\ , \tag{5.16}$$

$$F_2(t)=q+\sqrt{\frac{q(1-q)}{n}}\,v+\frac{1}{2}\,\frac{q(1-q)}{n}\,b_2v^2+o(1/n)\ , \tag{5.17}$$

where

$$c_{11} = \int_{-\infty}^{\eta} \frac{\partial f(\xi,y)}{\partial \xi}\, dy / f_1^2(\xi), \quad c_{12} = f(\xi,\eta)/f_1(\xi) f_2(\eta), \quad c_{22} = \int_{-\infty}^{\xi} \frac{\partial f(x,\eta)}{\partial \eta}\, dx / f_2^2(\eta),$$

$$b_1 = \frac{df_1(\xi)}{d\xi} / f_1^2(\xi) \; , \quad b_2 = \frac{df_2(\eta)}{d\eta} / f_2^2(\eta) \; .$$

Taking the logarithms of both sides of (5.6), inserting above expansions, and making the Taylor's expansions, we have, after some calculation,

$$\begin{aligned}
\log G_{2n} = {} & \sqrt{\frac{p(1-p)}{n}}\, u \left\{ \frac{n_1 h}{A_{11}} + \frac{n_2(1-h)}{A_{12}} - \frac{n_3 h}{A_{21}} - \frac{n_4(1-h)}{A_{22}} \right] \\
& + \sqrt{\frac{q(1-q)}{n}}\, v \left\{ \frac{n_1 h^*}{A_{11}} - \frac{n_2 h^*}{A_{12}} + \frac{n_3(1-h^*)}{A_{21}} - \frac{n_4(1-h^*)}{A_{22}} \right] \\
& + \frac{1}{2} \frac{p(1-p)}{n} u^2 \left\{ \frac{n_1 c_{11}}{A_{11}} + \frac{n_2(b_1 - c_{11})}{A_{12}} - \frac{n_3 c_{11}}{A_{21}} - \frac{n_4(b_1 - c_{11})}{A_{22}} - \right. \\
& \qquad \left. - \frac{n_1 h^2}{A_{11}^2} - \frac{n_2(1-h)^2}{A_{12}^2} - \frac{n_3 h^2}{A_{21}^2} - \frac{n_4(1-h)^2}{A_{22}^2} \right] \\
& + \frac{\sqrt{p(1-p)q(1-q)}}{n}\, uv \left\{ \frac{n_1 c_{12}}{A_{11}} - \frac{n_2 c_{12}}{A_{12}} - \frac{n_3 c_{12}}{A_{21}} + \frac{n_4 c_{12}}{A_{22}} - \right. \\
& \qquad \left. - \frac{n_1 h h^*}{A_{11}^2} + \frac{n_2(1-h)h^*}{A_{12}^2} + \frac{n_3 h(1-h^*)}{A_{21}^2} - \frac{n_4(1-h)(1-h^*)}{A_{22}^2} \right] \\
& + \frac{1}{2} \frac{q(1-q)}{n} v^2 \left\{ \frac{n_1 c_{22}}{A_{11}} - \frac{n_2 c_{22}}{A_{12}} + \frac{n_3(b_2 - c_{22})}{A_{21}} - \frac{n_4(b_2 - c_{22})}{A_{22}} - \right. \\
& \qquad \left. - \frac{n_1 h^{*2}}{A_{11}^2} - \frac{n_2 h^{*2}}{A_{12}^2} - \frac{n_3(1-h^*)^2}{A_{21}^2} - \frac{n_4(1-h^*)^2}{A_{22}^2} \right] \\
& + o(1) \; .
\end{aligned} \tag{5.18}$$

Now we change n_1, n_2, n_3, n_4 to w_1, w_2, w_3 by using (5.9) and by the relation $w_1+w_2+w_3 = -(n_4-nA_{22})/\sqrt{n} - 2/\sqrt{n}$ and then we use the relations (5.13) and (5.14). Then, keeping in mind that $|n_{11}-nA_{11}|n^{-1} \to 0$ in probability, (5.18) can be reduced to the form

$$(5.19) \qquad \log G_{2n} = (d_1u+d_2v)w_1 - \tfrac{1}{2}\{\ell_{11}u^2 - 2\ell_{12}uv + \ell_{22}v^2\} + o(1)$$

where

$$d_1 = \sqrt{p(1-p)}\,\{k^2h - (\frac{1}{A_{12}} + \frac{1}{A_{22}})\}\ ,$$

$$d_2 = \sqrt{q(1-q)}\,\{k^2h^* - (\frac{1}{A_{21}} + \frac{1}{A_{22}})\}\ ,$$

$$\ell_{11} = p(1-p)\{k^2h^2 - 2h(\frac{1}{A_{12}} + \frac{1}{A_{22}}) + (\frac{1}{A_{12}} + \frac{1}{A_{22}})\}\ ,$$

$$\ell_{12} = \sqrt{p(1-p)q(1-q)}\{-k^2hh^* + (\frac{1}{A_{21}} + \frac{1}{A_{22}})h + (\frac{1}{A_{12}} + \frac{1}{A_{22}})h^* - \frac{1}{A_{22}}\}\ ,$$

$$\ell_{22} = q(1-q)\{k^2h^{*2} - 2(\frac{1}{A_{21}} + \frac{1}{A_{22}})h^* + (\frac{1}{A_{21}} + \frac{1}{A_{22}})\}\ .$$

Thus, we have

$$(5.20) \qquad G_{2n} = \exp\{(d_1u+d_2v)w_1\}\exp\{-\tfrac{1}{2}(\ell_{11}u^2 - 2\ell_{12}uv + \ell_{22}v^2)\}\cdot\{1+o(1)\}$$

Summarizing (5.7), (5.12) and (5.20), (5.3) can be written as

$$\begin{aligned}
g_n'(u,v) &= hh^* \sqrt{p(1-p)q(1-q)}\,[A_{11}A_{12}A_{21}A_{22}(2\pi)^3]^{-\frac{1}{2}} \exp[-\tfrac{1}{2}(\ell_{11}u^2-2\ell_{12}uv+\ell_{22}v^2)] \times \\
&\quad \times \Sigma \exp[(d_1u+d_2v)w_1] \exp[-\tfrac{1}{2}k^2w_1^2]\Delta w_1 + o(1) \\
&= hh^* \sqrt{p(1-p)q(1-q)}\,[A_{11}A_{12}A_{21}A_{22}(2\pi)^3]^{-\frac{1}{2}} \exp[-[(\ell_{11}-\frac{d_1^2}{k^2})u^2 - \\
&\quad -2(\ell_{12}+\frac{d_1d_2}{k^2})vu + (\ell_{22}-\frac{d_2^2}{k^2})v^2]] \times \qquad (5.21) \\
&\quad \times \Sigma \exp[-\tfrac{1}{2}k^2(w_1-\frac{d_1u+d_2v}{k^2})^2]\Delta w_1 + o(1), \\
&= hh^* \sqrt{p(1-p)q(1-q)}\,[A_{11}A_{12}A_{21}A_{22}(2\pi)^3]^{-\frac{1}{2}} \exp[-\tfrac{1}{2}\frac{p(1-p)q(1-q)}{k^2\,A_{11}A_{12}A_{21}A_{22}} \\
&\quad \cdot [u^2-2\frac{A_{11}A_{22}-A_{12}A_{21}}{\sqrt{p(1-p)q(1-q)}}uv+v^2]] \times \\
&\quad \times \Sigma \exp[-\tfrac{1}{2}k^2(w_1-\frac{d_1u+d_2v}{k^2})^2]\Delta w_1 + o(1)
\end{aligned}$$

The summation in the above expression is taken over the discrete values of w_1 corresponding to the values of n_{11} in the interval $[-\sqrt{n}\,A_{11}, \sqrt{n}\min(A_{12},A_{21}) + O(\frac{1}{\sqrt{n}})]$. Since the summand is integrable in any interval in the real line with respect to w_1, then for large n, this summation can be approximated, to within terms of order $1/\sqrt{n}$, by

$$\int_{-\infty}^{\infty} \exp[-\tfrac{1}{2}k^2(w_1-\frac{d_1u+d_2v}{k^2})^2]dw_1$$

i.e., $\sqrt{2\pi}/k$ and tends to this value as its limit. Consequently, (5.21) becomes

$$g_n'(u,v) = hh^* \frac{1}{2\pi}\frac{1}{k} \sqrt{\frac{p(1-p)q(1-q)}{A_{11}A_{12}A_{21}A_{22}}} \cdot \exp[-\frac{1}{2}\frac{p(1-p)q(1-q)}{k^2 A_{11}A_{12}A_{21}A_{22}} \tag{5.22}$$
$$\cdot (u^2 - 2\,\frac{A_{11}A_{22}-A_{12}A_{21}}{\sqrt{p(1-p)q(1-q)}}\,uv + v^2)] + o(1)$$

If we put

$$\rho = \frac{A_{11}A_{22}-A_{12}A_{21}}{\sqrt{p(1-p)q(1-q)}} = \frac{F(\xi,\eta)-pq}{\sqrt{p(1-p)q(1-q)}} \quad , \tag{5.23}$$

the leading term of (5.22) has the form,

$$\lim_{n\to\infty} g_n'(u,v) = (hh^*)\cdot \frac{1}{2\pi\sqrt{1-\rho^2}} \exp[-\frac{1}{2(1-\rho^2)}(u^2-2\rho uv+v^2)] \ . \tag{5.24}$$

This is the limiting form of (2.10), in which only the factor hh^* depends on the location of two sample points determining our combined quantile, and it is easy to see that other three terms, i.e., (2.11), (2.12), and (2.13) can be reduced to limiting forms of the standard bivariate normal density times $h(1-h^*)$, $(1-h)h^*$ respectively. Since $hh^* + h(1-h^*)+(1-h)h^* + (1-h)(1-h^*) = 1$, we finally have for the full expression function, $g_n(u,v)$ of the normalized combined quantile, (u,v),

$$\lim_{n\to\infty} g_n(u,v) = g(u,v) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp[-\frac{1}{2(1-\rho^2)}(u^2-2\rho uv+v^2)]. \tag{5.25}$$

Since $g_n(u,v)$ is uniformly bounded in any two dimensional interval in (u,v)-space,

$$\lim_{n\to\infty} \int_a^b \int_c^d g_n(u,v)dudv = \int_a^b \int_c^d g(u,v)dudv \ . \tag{5.26}$$

Let us summarize the above argument in the

Theorem 5.1. *Let* (ξ,η) *be the population combined quantile of order* (p,q) $(0 < p,\ q < 1)$ *in the bivariate population of continuous type with c.d.f.* $F(x,y)$ *and d.f.* $f(x,y)$, *and let* $(z^{(1)},z^{(2)})$ *be the corresponding quantity in samples of size* n *taken from this population. Suppose that, in some neighborhood of the point* (ξ,η), $f(x,y)$ *are continuous and has continuous derivatives as a function of two variables, and* $f_1(\xi) \neq 0$ *and* $f_2(\eta) \neq 0$. *Then*

$$(u,v) = \left\{ \sqrt{\frac{n}{p(1-p)}}\, f_1(\xi)\ (z^{(1)}-\xi),\ \sqrt{\frac{n}{q(1-q)}}\, f_2(\eta)(z^{(2)}-\eta)\right\}$$

has a limiting standard bivariate normal distribution with correlation coefficient

$$\rho = \frac{F(\xi,\eta)-pq}{\sqrt{p(1-p)q(1-q)}} .$$

where $f_1(x)$ *and* $f_2(y)$ *are marginal d.f.'s of* x *and* y, *respectively*.

In the particular case when $p = q = 1/2$, i.e., in the case of the bivariate median $(\tilde{x},\tilde{y})$,

$$(u,v) = (2\sqrt{n}\ f_1(\xi)(\tilde{x}-\xi),\ \ 2\sqrt{n}\ f_2(\eta)(\tilde{y}-\eta))$$

and

$$\rho = 4(F(\xi,\eta)- \tfrac{1}{4}) .$$

If x and y are independent, ρ is zero.

Next we shall go to treat the general case. Let

$$(5.27) \qquad 0 < p_1 < p_2 <\cdots < p_{k_1} < 1 \ ;\ 0 < q_1 < q_2 <\cdots < q_{k_2} < 1$$

and

$$(5.28) \qquad \alpha_i = [np_i],\ i=1,\dots,k_1\ ;\ \beta_j = [np_j],\ j=1,\dots,k_2,$$

where $[\cdot\cdot]$ is the Gauss symbol and $k_1 \geqq k_2$. Since we are treating the case of large n, we may assume that $k_1 + k_2 < n$. Let us denote the population quantiles of order p_i of the x-marginal distribution and of order q_j of the y-marginal distribution by ξ_i and η_j, respectively, that is,

$$(5.29) \qquad F_1(\xi_i) = p_i,\ i=1,\dots,k\ ;\ F_2(\eta_j) = q_j,\ i=1,\dots,k_2\ .$$

Corresponding sample quantities are denoted by

$$z_1^{(1)},\ z_2^{(1)},\ \dots,\ {}^{(1)}_{k_1}$$

for the x-component and by

$$z_1^{(2)},\ z_2^{(2)},\dots,\ {}^{(2)}_{k_2}$$

for the y-component, which are, by definition, the (α_i+1)th order statistics in x-component, x_{α_i+1}, and the (β_j+1)th order statistics in y-component, y_{β_j+1}, respectively, where we assume that np_i and nq_j are not integral. Let us assume that in some neighborhood of points (ξ_i,η_j), $i=1,\dots,k_1$; $j=1,\dots,k_2$, $f(x,y)$ is continuous and has continuous derivatives as a function of two variables, and $f_1(\xi_i) \neq 0$, $i=1,\dots,k_1$ and $f_2(\eta_j) \neq 0$, $j=1,\dots,k_2$.

Since $k_1 + k_2 \leqq n$, the exact joint density function of the $z_i^{(1)}$'s and the $z_j^{(2)}$'s is, as seen in Subsection 3.4, expressed as the sum of k_2+1 contributions due to the cases distinguished

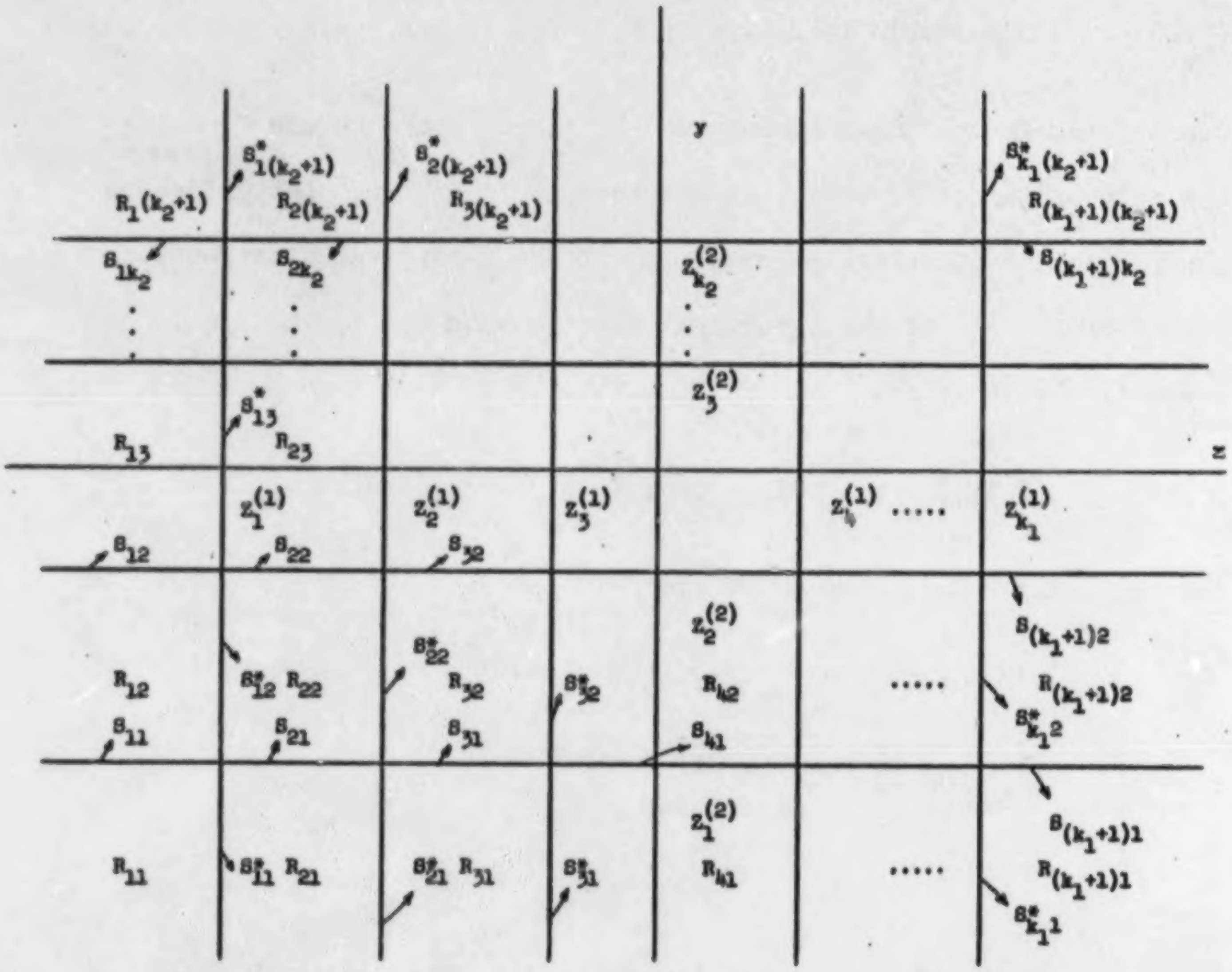

y
$S^*_{1(k_2+1)}$
$S^*_{2(k_2+1)}$
$S^*_{k_1(k_2+1)}$
$R_{1(k_2+1)}$
$R_{2(k_2+1)}$
$R_{3(k_2+1)}$
$R_{(k_1+1)(k_2+1)}$
S_{1k_2}
S_{2k_2}
$Z^{(2)}_{k_2}$
$S_{(k_1+1)k_2}$
$Z^{(2)}_3$
S^*_{13}
R_{13}
R_{23}
z
$Z^{(1)}_1$
$Z^{(1)}_2$
$Z^{(1)}_3$
$Z^{(1)}_4$ .....
$Z^{(1)}_{k_1}$
S_{12}
S_{22}
S_{32}
$Z^{(2)}_2$
$S_{(k_1+1)2}$
S^*_{22}
R_{12}
S^*_{12} R_{22}
R_{32}
S^*_{32}
R_{42}
.....
$R_{(k_1+1)2}$
$S^*_{k_1 2}$
S_{11}
S_{21}
S_{31}
S_{41}
$Z^{(2)}_1$
$S_{(k_1+1)1}$
R_{11}
S^*_{11} R_{21}
S^*_{21} R_{31}
S^*_{31}
R_{41}
.....
$R_{(k_1+1)1}$
$S^*_{k_1 1}$

Figure 5. Quantiles $Z^{(1)}_1,\ldots,Z^{(1)}_{k_1}$ and $Z^{(2)}_1,\ldots,Z^{(2)}_{k_2}$ and notations used in the discussion.

by the number of sample points determining our order statistics. As in the case of a single combined quantile, it is to be noted that, among these contributions, the one due to the last extreme case, i.e., the case when $z_i^{(1)}$'s and $z_j^{(2)}$'s are determined by $k_1 + k_2$ different sample points, is of the largest order with respect to n and others are of order $1/n$, $1/n^2, \ldots$ in comparison with it, and hence may be neglected in the asymptotic joint distribution of $z_i^{(1)}$'s and $z_j^{(2)}$'s. In the extreme case to be considered, there are $(k_1+1)^{k_2}(k_2+1)^{k_1}$ minor cases of same order of n according to ways in which strip-portions k_1+k_2 sample points determining $z_i^{(1)}$'s and $z_j^{(2)}$'s lie (see Figure 5).

Now let us pick up one of these minor cases, for example, the case when k_1+k_2 sample points lie in strips, $S_{11}, S_{12}, \ldots, S_{1k_2}$; $S^*_{11}, S^*_{21}, \ldots, S^*_{k_1 1}$ and evaluate the contribution due to this particular case when n is large. Let us use the following notations with the symbols in Figure 5:

$$a_{ij} = \int\!\!\int_{R_{ij}} f(x,y)dxdy, \quad i=1,\ldots,k_1+1;\ j=1,\ldots,k_2+1,$$

$$\bar{h}_i = \int_{S^*_{i1}} f(z_i^{(1)},y)dy, \ i=1,\ldots,k_1; \ \bar{h}_j^* = \int_{S_{1j}} f(x,z_j^{(2)})dx, \quad j=1,\ldots,k_2$$

n_{ij} = the number of sample points falling in the region R_{ij}.

Then the contribution denoted by $g_n^*(z^{(1)},z^{(2)})$, can be written in the form

$$(5.30) \qquad g_n^*(z^{(1)},z^{(2)}) = \prod_{i=1}^{k_1} \bar{h}_i \prod_{j=1}^{k_2} \bar{h}_j^* \sum \frac{n!}{\prod_{i=1}^{k_1+1}\prod_{j=1}^{k_2+1}(n_{ij})!} \prod_{i=1}^{k_1+1}\prod_{j=1}^{k_2+1} a_{ij}^{n_{ij}},$$

where the summation is taken over the range of n_{ij} satisfying the constraints

$$(5.31)\qquad \sum_{j=1}^{k_2+1} n_{ij} \equiv N_{i\cdot} = \begin{cases} \alpha_1-k_2 & \text{for } i=1,\\ \alpha_i-\alpha_{i-1}-1 & \text{for } i=2,\dots,k_1,\\ n-\alpha_{k_1}-1 & \text{for } i=k_1+1\ , \end{cases}$$

$$(5.32)\qquad \sum_{i=1}^{k_1+1} n_{ij} \equiv N_{\cdot j} = \begin{cases} \beta_1-k_1 & \text{for } j=1,\\ \beta_j-\beta_{j-1}-1 & \text{for } j=2,\dots,k_2\\ n-\beta_{k_2}-1 & \text{for } j=k_2+1 \end{cases}$$

$$(5.33)\qquad \sum_{i=1}^{k_1+1}\ \sum_{j=1}^{k_2+1} n_{ij} = n-(k_1+k_2)\ .$$

It should be noted that the number of independent variables among n_{ij}'s is

$$(5.34)\qquad (k_1+1)(k_2+1)-(k_1+1)-(k_2+1)+1 = k_1k_2\ .$$

Now we introduce the new variables

$$(5.35)\qquad u_i = \sqrt{\frac{n}{p_i(1-p_i)}}\, f_1(\xi_i)(z_i^{(1)}-\xi_i)\ ,\ i=1,2,\dots,k_1,$$

$$(5.36)\qquad v_j = \sqrt{\frac{n}{q_j(1-q_j)}}\, f_2(\eta_j)(z_j^{(2)}-\eta_j)\ ,\ j=1,2,\dots,k_2.$$

Then $g_n^*(z^{(1)},z^{(2)})$ is reduced to

$$(5.37)\qquad g_n'(u,v) = \prod_{i=1}^{k_1} \frac{\sqrt{p_i(1-p_i)}}{f_1(\xi_i)} \prod_{j=1}^{k_2} \frac{\sqrt{q_j(1-q_j)}}{f_2(\eta_j)}\ \frac{1}{n^{\frac{1}{2}(k_1+k_2)}}\, g_n^*(u,v),$$

where $g_n^*(u,v)$ is a function of u_i's and v_j's obtained from $g_n^*(z^{(1)},z^{(2)})$ by just putting

$$(5.35') \qquad z_i^{(1)} = \xi_i + \sqrt{\frac{p_i(1-p_i)}{n}}\ \frac{1}{f_1(\xi_i)}\, u_i\ , \quad i=1,\dots,k_1\ ,$$

$$(5.36') \qquad z_j^{(2)} = \eta_j + \sqrt{\frac{q_j(1-q_j)}{n}}\ \frac{1}{f_2(\eta_j)}\, v_j\ , \quad j=1,\dots,k_2\ .$$

In order to evaluate the behavior of $g_n'(u,v)$ for large n, we shall divide it into four factors, i.e.,

$$(5.38) \qquad g_n'(u,v) = \prod_{i=1}^{k_1} \sqrt{p_i(1-p_i)}\ \prod_{j=1}^{k_2} \sqrt{q_j(1-q_j)}\ G_{on} \sum G_{1n}G_{2n},$$

where

$$(5.39) \qquad G_{on} = \prod_{i=1}^{k_1} \frac{\bar{h}_i}{f_1(\xi_i)}\ \prod_{j=1}^{k_2} \frac{\bar{h}_j^*}{f_2(\eta_n)}\ ,$$

$$(5.40) \qquad G_{1n} = \frac{1}{n^{\frac{1}{2}(k_1+k_2)}}\ \frac{n!}{\prod\limits_{i=1}^{k_1+1} \prod\limits_{j=1}^{k_2+1} (n_{ij})!}\ \prod_{i=1}^{k_1+1} \prod_{j=1}^{k_2+1} A_{ij}^{n_{ij}}\ ,$$

$$(5.41) \qquad G_{2n} = \prod_{i=1}^{k_1+1} \prod_{j=1}^{k_2+1} \left(\frac{a_{ij}}{A_{ij}}\right)^{n_{ij}}\ ,$$

where A_{ij} is the corresponding probability to a_{ij} when $z_i^{(1)} = \xi_i$ and $z_j^{(2)} = \eta_j$. First of all, we easily have

$$(5.42) \qquad G_{on} = \prod_{i=1}^{k_1} h_i \prod_{j=1}^{k_2} h_j^* + O\left(\frac{1}{\sqrt{n}}\right)\ ,$$

where

$$(5.43) \qquad h_i = \int_{-\infty}^{\eta_1} f(\xi_i, y)\,dy / f_1(\xi_i), \quad i=1,\ldots,k_1 ,$$

$$(5.44) \qquad h_j^* = \int_{-\infty}^{\xi_1} f(x, \eta_j)\,dx / f_2(\eta_j), \quad j=1,\ldots,k_2 .$$

Secondly

$$G_{1n} = n^{\frac{1}{2}(k_1+k_2)} \left\{ \frac{(n-k_1-k_2)!}{\prod_{i=1}^{k_1+1} \prod_{j=1}^{k_2+1} (n_{ij})!} \prod_{i=1}^{k_1+1} \prod_{j=1}^{k_2+1} A_{ij}^{n_{ij}} \right\} [1+O(1/n)] .$$

For the factor in the curled parentheses, we can use the well-known normal approximation for the multinomial distribution. According to the law of large numbers, we have only to pay our attention to the values of n_{ij} such that $|n_{ij}-nA_{ij}|n^{-1} \to 0$. Let us introduce the variables

$$(5.45) \qquad w_{ij}^* = (n_{ij}-nA_{ij})/\sqrt{n} , \quad i=1,\ldots,k_1+1 ; \; j=1,\ldots,k_2+1 ,$$

and renumber them in a certain order, for example, $w_1^* = w_{11}^*, \ldots, w_{k_2+1}^* = w_{1,k_2+1}^*$, $w_{k_2+2}^* = w_{21}^*, \ldots, w_{2(k_2+1)}^* = w_{2,k_2+1}^*, \ldots, w_{k_1(k_2+1)+1}^* = w_{k_1+1,1}^*, \ldots,$ $w_{(k_1+1)(k_2+1)}^* = w_{(k_1+1),(k_2+1)}^*$, and put $k = k_1+k_2$, $d = k_1 \cdot k_2$ and $r=(k_1+1)(k_2+1)$. Then we have

$$(5.46) \qquad G_{1n} = n^{k/2}[|M|/(2\pi)^{r-1}]^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\sum_{i,j=1}^{r-1} m_{ij} w_i^* w_j^*\right\} \prod_{i=1}^{r-1} \Delta w_i^* + O(1/\sqrt{n}),$$

where

$$\Delta w_i^* = 1/\sqrt{n} , \quad i=1,2,\ldots,r-1 ,$$

$$m_{ij} = 1/A_{(k_1+1),(k_2+1)} \ , \ i \neq j$$

$$m_{ii} = \begin{cases} 1/A_{1i} + 1/A_{(k_1+1),(k_2+1)} \ , & i=1,\ldots,k_2+1 \\ \cdots\cdots\cdots\cdots & \cdots\cdots \\ 1/A_{k_1 i} + 1/A_{(k_1+1),(k_2+1)} \ , & i=k_1(k_2+1)+1,\ldots,r-1, \end{cases}$$

and $|M|$ is the determinant of $(r-1) \times (r-1)$ matrix (m_{ij}), which is $\prod_{i=1}^{k_1+1} \prod_{j=1}^{k_2+1} 1/A_{ij}$. However, on account of the linear constrains on n_{ij}'s, every w_i^* can be expressed by d independent w_i^*'s, say, $w_1, w_2, \ldots, w_d$, in the form

$$w_i^* = c_{i1}w_1 + c_{i2}w_2 + \cdots + c_{id}w_d + O(1/\sqrt{n}) \ , \ i=1,\ldots,r-1$$

or in the matrix notation

$$(5.47) \qquad w^* = C\, w + \underline{O}(1/\sqrt{n}),$$

where $w^{*\prime} = (w_1^*,\ldots,w_{r-1}^*)$, $w' = (w_1,\ldots,w_d)$, $C = (c_{ij})$ is $(r-1) \times d$ matrix and obviously is of rank d, and $\underline{O}(1/\sqrt{n})$ is a $(r-1) \times 1$ vector in which d elements are zero and other $r-d-1$ elements are of order $1/\sqrt{n}$. Substituting (5.47) into (5.46), G_{1n} can be expressed in the form

$$(5.48) \qquad G_{1n} = [|M|/(2\pi)^{r-1}]^{1/2} \exp\left\{-\frac{1}{2} w' B w\right\} \prod_{i=1}^{d} \Delta w_i + O(1/\sqrt{n}),$$

where

$$\Delta w_i = 1/\sqrt{n} \ , \ i=1,\ldots,d \ ,$$

and

and

$$B = C'(m_{ij})C \tag{5.49}$$

is a symmetric positive definite matrix.

Finally we shall treat G_{2n}. To do so, we express every a_{ij} by d.f. as

$$a_{ij} = F(z_i^{(1)},z_j^{(2)})-F(z_i^{(1)},z_{j-1}^{(2)})-F(z_{i-1}^{(1)},z_j^{(2)})+F(z_{i-1}^{(1)},z_{j-1}^{(2)}), \tag{5.50}$$

with the following conventions:

$$F(z_i^{(1)},z_o^{(2)}) = F(z_o^{(1)},z_j^{(2)}) = 0 \ , \ F(z_{k_1+1}^{(1)},z_j^{(2)}) = F_2(z_j^{(2)}) \ ,$$

$$F(z_i^{(1)},z_{k_2+1}^{(2)}) = F_1(z_i^{(1)})$$

$$F(z_{k_1+1}^{(1)},z_{k_2+1}^{(2)}) = 1 \ ,$$

and A_{ij} by d.f. with arguments ξ_i and η_j in a similar way. Using the Taylor expansion like (5.15) for each $F(z_i^{(1)},z_j^{(2)})$ about (ξ_i,η_j), we have

$$\begin{aligned}
\log G_{2n} &= \sum_{i=1}^{k_1+1} \sum_{j=1}^{k_2+1} n_{ij}\log(\frac{a_{ij}}{A_{ij}}) \\
&= \sum_{i=1}^{k_1+1} \sum_{j=1}^{k_2+1} (\sqrt{n}\, w_{ij}^{*}+nA_{ij})\Big[\frac{1}{\sqrt{n}}\frac{1}{A_{ij}}\Big\{\ell_{ij}u_i-\ell_{(i-1)j}u_{i-1}+f_{ij}v_j-f_{i(j-1)}v_{j-1}\Big\} \\
&\quad + \frac{1}{2n}\frac{1}{A_{ij}}\Big\{(\lambda_{ij}-\frac{\ell_{ij}^2}{A_{ij}})u_i^2+2\frac{\ell_{ij}\ell_{(i-1)j}}{A_{ij}}u_iu_{i-1}-(\lambda_{(i-1)j}+\frac{\ell_{(i-1)j}^2}{A_{ij}})u_{i-1}^2\Big\} \\
&\quad + \frac{1}{2n}\frac{1}{A_{ij}}\Big\{(\mu_{ij}-\frac{\ell_{ij}f_{ij}}{A_{ij}})u_iv_j-2(\mu_{i(j-1)}-\frac{\ell_{ij}f_{i(j-1)}}{A_{ij}})u_iv_{j-1}
\end{aligned}$$

(5.51)

$$-2\left(\mu_{(i-1)j}-\frac{\ell_{(i-1)j}f_{ij}}{A_{ij}}\right)u_{i-1}v_j+2\left(\mu_{(i-1)(j-1)}-\frac{\ell_{(i-1)j}f_{i(j-1)}}{A_{ij}}\right)u_{i-1}v_{j-1}\Big\}$$

$$+\frac{1}{2n}\frac{1}{A_{ij}}\left\{\left(\nu_{ij}-\frac{f_{ij}^2}{A_{ij}}\right)v_j^2+2\,\frac{f_{ij}f_{i(j-1)}}{A_{ij}}\,v_jv_{j-1}-\left(\nu_{i(j-1)}+\frac{f_{i(j-1)}^2}{A_{ij}}\right)v_{j-1}^2\right\}$$

$$+\;o(1/n)\Big]\quad ,$$

where

$$\ell_{ij}=\sqrt{p_i(1-p_i)}\int_{\eta_{j-1}}^{\eta_j} f(\xi_i,y)dy/f_1(\xi_i),\quad f_{ij}=\sqrt{q_j(1-q_j)}\int_{\xi_{i-1}}^{\xi_i} f(x,\eta_j)dx/f_2(\eta_j),$$

$$\lambda_{ij}=p_i(1-p_i)\int_{\eta_{j-1}}^{\eta_j}\frac{\partial f(\xi_i,y)}{\partial\xi_i}\,dy/f_1^2(\xi_i),\quad \mu_{ij}=\sqrt{p_i(1-p_i)q_j(1-q_j)}\;\times$$

$$\times\; f(\xi_i,\eta_j)/f_1(\xi_i)f_2(\eta_j),$$

$$\nu_{ij}=q_j(1-q_j)\int_{\xi_{i-1}}^{\xi_i}\frac{\partial f(x,\eta_j)}{\partial\eta_j}\,dx/f_2^2(\eta_j)\quad .$$

Since, for fixed u_i's and v_j's, $w_{ij}^*/\sqrt{n}\to 0$ in probability and since

$$\sum_{i=1}^{k_1+1}\{\ell_{ij}u_i-\ell_{(i-1)j}u_{i-1}\}=0\ ,\quad \sum_{j=1}^{k_2+1}\{f_{ij}v_j-f_{i(j-1)}v_{j-1}\}=0\ ,$$

$$\sum_{i=1}^{k_1+1}\{\lambda_{ij}u_i^2-\lambda_{(i-1)j}u_{i-1}^2\}=0\ ,\quad \sum_{j=1}^{k_2+1}\{\nu_{ij}v_j^2-\nu_{i(j-1)}v_{j-1}^2\}=0\ ,$$

$$\sum_{i=1}^{k_1+1}\{\mu_{ij}u_i-\mu_{(i-1)j}u_{i-1}\}=0\ ,\quad \sum_{i=1}^{k_1+1}\{\mu_{i(j-1)}u_i-\mu_{(i-1)(j-1)}u_{i-1}\}=0,$$

then

$$\begin{aligned}\log G_{2n} = \sum_{i=1}^{k_1+1} \sum_{j=1}^{k_2+1} [& \{\ell_{ij}u_i - \ell_{(i-1)j}u_{i-1} + f_{ij}v_j - f_{i(j-1)}v_{j-1}\} \frac{w^*_{ij}}{A_{ij}} \\ & - \frac{1}{2A_{ij}} \{\ell^2_{ij}u^2_i - 2\ell_{ij}\ell_{(i-1)j}u_iu_{i-1} + \ell^2_{(i-1)j}u^2_{i-1} - 2\ell_{ij}f_{ij}u_iv_j + \\ & \qquad + 2\ell_{ij}f_{i(j-1)}u_iv_{j-1} \\ & + 2\ell_{(i-1)j}f_{ij}u_{i-1}v_j - 2\ell_{(i-1)j}f_{i(j-1)}u_{i-1}v_{j-1} + f^2_{ij}v^2_j - \\ & \qquad - 2f_{ij}f_{i(j-1)}v_jv_{j-1} + f^2_{i(j-1)}v^2_{j-1}\} \\ & + o(1)] \\ = \sum_{i=1}^{k_1+1} \sum_{j=1}^{k_2+1} & \{\ell_{ij}u_i - \ell_{(i-1)j}u_{i-1} + f_{ij}v_j - f_{i(j-1)}v_{j-1}\} \frac{w^*_{ij}}{A_{ij}} \\ & - \frac{1}{2} \sum_{i=1}^{k_1+1} \sum_{j=1}^{k_2+1} \frac{1}{A_{ij}} \{\ell_{ij}u_i - \ell_{(i-1)j}u_{i-1} - f_{ij}v_j + f_{i(j-1)}v_{j-1}\}^2 \\ & + o(1) \ .\end{aligned} \tag{5.52}$$

If we express w^*_{ij} by independent variables $w_1, \ldots, w_\alpha$, the first sum in (5.52) can be written as

$$\sum_{\alpha=1}^{d} \ell_\alpha(u,v)w_\alpha = \ell'w \ , \tag{5.53}$$

where $\ell_\alpha(u,v) \equiv \ell_\alpha$ are linear expressions of $u' = (u_1, \ldots, u_{k_1})$ and $v' = (v_1, \ldots, v_{k_2})$, i.e., with the $d \times k$ coefficient matrix D, $\ell' = (\ell_1, \ldots, \ell_d)$ can be expressed by

$$\ell = D[{}^{u}_{v}] \ . \tag{5.54}$$

The second sum in (5.52) is a positive definite quadratic form of u and v,

$$[u',v']\, G\left[\begin{smallmatrix}u\\v\end{smallmatrix}\right] \tag{5.55}$$

where G is the $k \times k$ positive definite matrix. Consequently, we have

$$G_{2n} = \exp(\ell' w)\exp\{-\tfrac{1}{2}[u',v']\, G\left[\begin{smallmatrix}u\\v\end{smallmatrix}\right]\} + o(1)\ . \tag{5.56}$$

Combining (5.42), (5.48) and (5.56) together, (5.58) can be written as

$$\begin{aligned} g_n'(u,v) = {} & \prod_{i=1}^{k_1} h_i \prod_{j=1}^{k_2} h_j^* \prod_{i=1}^{k_1} \sqrt{p_i(1-p_i)} \prod_{j=1}^{k_2} \sqrt{q_j(1-q_j)}\ \{|M|/(2\pi)^{r-1}\}^{\frac{1}{2}} \cdot \\ & \cdot \exp\{-\tfrac{1}{2}[u',v']\, G\left[\begin{smallmatrix}u\\v\end{smallmatrix}\right]\} \times \sum \exp(\ell' w)\cdot\exp(-\tfrac{1}{2} w' B w)\cdot \\ & \cdot \prod_{i=1}^{d} \Delta w_i + o(1)\ , \end{aligned} \tag{5.57}$$

where the summation is taken over a range of $w_1,\ldots,w_\alpha$ which is tending to the whole d-mensional euclidean space R_d as $n \to \infty$. Since $\exp(\ell' w)\cdot\exp(-\frac{1}{2} w' B w)$ is integrable in R_d, we can carry out the summation, to within terms of order $o(1)$, by the integration with respect to $w_1,\ldots,w_\alpha$ over R_d. Then, since

$$\int\!\cdots\!\int_{R_d} \exp\{-\tfrac{1}{2} w' B w + \ell' w\} dw_1 \cdots dw_d = \{(2\pi)^d/|B|\}^{\frac{1}{2}} \exp\{\tfrac{1}{2}\ell' B^{-1}\ell\},$$

we have

$$g_n'(u,v) = \left(\prod_{i=1}^{k_1} h_i \prod_{j=1}^{k_2} h_j^*\right) K \exp\{-\tfrac{1}{2}[u',v']G[{}^u_v]\}\cdot\exp\{\tfrac{1}{2}\, \ell' B^{-1}\ell\} + o(1)$$

(5.58)
$$= \left(\prod_{i=1}^{k_1} h_i \prod_{j=1}^{k_2} h_j^*\right) K \exp\{-\tfrac{1}{2}[u',v'](G-D'B^{-1}D)[{}^u_v]\} + o(1) \;,$$

where

(5.59)
$$K = \prod_{i=1}^{k_1} \sqrt{p_i(1-p_i)} \prod_{j=1}^{k_2} \sqrt{q_j(1-q_j)} \left\{\frac{|M|}{|B|}\right\}^{\frac{1}{2}}(2\pi)^{-k/2} \;.$$

This is the expression for our present particular case when $k_1 + k_2$ sample points determining the $z_i^{(1)}$'s and $z_j^{(2)}$'s lie in the strips $S_{11}, S_{12}, \ldots, S_{1k_2}$; $S_{11}^*, S_{21}^*, \ldots, S_{k_1 1}^*$. In this, only the factor $\left(\prod_{i=1}^{k_1} h_i \prod_{j=1}^{k_2} h_j^*\right)$ depends on this particularity. Other cases have the corresponding factors. Since it is easy to see that the sum of all these factors is equal to one, we have for the full density function of u and v, which is denoted by $g_n(u,v)$,

(5.60)
$$\lim_{n\to\infty} g_n(u,v) = g(u,v) = K \exp\{-\tfrac{1}{2}[u',v']\,\Lambda^{-1}[{}^u_v]\}$$

where

(5.61)
$$\Lambda^{-1} = G-D'B^{-1}D \;,$$

and at this stage, this does not mean the inverse of Λ, but is just an abbreviated symbol of the $k \times k$ symmetric matrix in the right-hand side. Since u and v are arbitrary, this limit exists almost everywhere in R_k. If we can evaluate the Λ^{-1} concretely, the final result will be obtained. This work is, however, terribly tedious one. Fortunately we can avoid it on account of the result obtained for the case of single combined quantile and of the Mosteller theorem [5] for the univariate case.

Since $g_n(u,v)$ is the density and is finite almost everywhere for every n, $g_n(u,v)$ is uniformly bounded in any k-dimensional interval I_k in R_k. Hence

$$\int_{I_k}\cdots\int g_n(u,v)dudv \to \int_{I_k}\cdots\int g(u,v)dudv \leqq 1. \tag{5.62}$$

From this, Λ^{-1} and hence Λ should be positive definite (now Λ^{-1} is the inverse of Λ), and since, for $I_k = R_k$, the equality in (5.62) holds, $K = (2\pi)^{-k/2}|\Lambda|^{-1/2}$. Thus, we have

$$g(u,v) = (2\pi)^{-k/2}|\Lambda|^{-1/2}\exp\{-\tfrac{1}{2}[u',v']\,\Lambda^{-1}[\begin{smallmatrix}u\\ v\end{smallmatrix}]\} . \tag{5.63}$$

This is a k-dimensional normal density function, and hence Λ is the covariance matrix in this limiting distribution. If we denote Λ as

$$\Lambda = \begin{bmatrix}\Lambda_{uu} & \Lambda_{uv}\\ \Lambda_{vu} & \Lambda_{vv}\end{bmatrix} , \tag{5.64}$$

where $\Lambda_{uu} = (\phi_{ij}^{(uu)})$ is the $k_1 \times k_1$ covariance matrix of u, $\Lambda_{vv} = (\phi_{ij}^{(vv)})$ is the $k_2 \times k_2$ covariance matrix of v, and $\Lambda_{uv} = (\phi_{ij}^{(uv)}) = \Lambda'_{vu}$ is the $k_1 \times k_2$ covariance matrix between u and v. From Mosteller's theorem,

$$\phi_{ij}^{(uu)} = \text{cov}(u_i,u_j) = \sqrt{p_i(1-p_j)/p_j(1-p_i)} \text{ for } i \leqq j ; \tag{5.65}$$

$$= \sqrt{p_j(1-p_i)/p_i(1-p_j)} \text{ for } i \geqq j ,$$

and

$$(5.66)\qquad \phi_{ij}^{(vv)} = \operatorname{cov}(v_i, v_j) = \sqrt{q_i(1-q_j)/q_j(1-q_i)} \quad \text{for } i \leqq j\ ;$$

$$= \sqrt{q_j(1-q_i)/q_i(1-q_j)} \quad \text{for } i \geqq j\ ,$$

and from the result of our Theorem 5.1,

$$(5.67)\qquad \phi_{ij}^{(uv)} = \operatorname{cov}(u_i, v_j) = \frac{F(\xi_i, \eta_j) - p_i q_j}{\sqrt{p_i(1-p_i)q_j(1-q_j)}}\ ,$$

where 'cov' means covariance in the limiting distribution.

To summarize:

Theorem 5.2. *Suppose that we have a bivariate population of continuous type with c.d.f*. $F(x,y)$, *d.f*. $f(x,y)$, *and with marginal c.d.f.'s*, $F_1(x)$, $F_2(y)$ *and marginal d.f.'s* $f_1(x)$, $f_2(y)$. *Let* $\xi_1, \ldots, \xi_{k_1}$ *be* k_1 *population quantiles of order* $0 < p_1 < \cdots < p_{k_1} < 1$ *in* $F_1(x)$, *i.e*., $F_1(\xi_i) = p_i$, $i=1,\ldots,k_1$ *and let* $\eta_1, \ldots, \eta_{k_2}$ *be* k_2 *population quantiles of order* $0 < q_1 < \cdots < q_{k_2} < 1$ *in* $F_2(y)$, *i.e*., $F_2(\eta_j) = q_j$, $j=1,\ldots,k_2$. *Let* $z_1^{(1)}, \ldots, z_{k_1}^{(1)}$ *for x-component and* $z_1^{(2)}, \ldots, z_{k_2}^{(2)}$ *for y-component be corresponding quantities in the sample of size* n *taken from this population, that is,* $z_i^{(1)}$ *is the* $[np_i]+1^{th}$ *order statistic in* x *and* $z_j^{(2)}$ *is the* $[nq_j]+1^{th}$ *order statistic in* y. *Suppose that, in some neighborhood of* (ξ_i, η_j), $i=1,\ldots,k_1$; $j=1,\ldots,k_2$, $f(x,y)$ *is continuous and has continuous derivatives as a function of two variables, and* $f_1(\xi_i) \neq 0$, $f_2(\eta_j) \neq 0$, $i=1,\ldots,k_1$; $j=1,\ldots,k_2$. *Then, the joint distribution of* $z_1^{(1)}, \ldots, z_{k_1}^{(1)}$ *and* $z_1^{(2)}, \ldots, z_{k_2}^{(2)}$ *tends to a* $k=k_1+k_2$ *dimensionsl normal distribution with means* $\xi_1, \ldots, \xi_{k_1}$; $\eta_1, \ldots, \eta_{k_2}$ *and with covariance*

$$(5.68) \qquad \frac{1}{n}\Sigma = \frac{1}{n}\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

where

$$(5.69) \qquad \underset{k_1\times k_1}{\Sigma_{11}} = (\sigma_{ij}^{(11)}) \ , \ \sigma_{ij}^{(11)} = p_i(1-p_j)/f_1(\xi_i)f_1(\xi_j) \quad \text{for } i \leqq j$$

$$= p_j(1-p_i)/f_1(\xi_i)f_1(\xi_j) \quad \text{for } i \geqq j$$

$$(5.70) \qquad \underset{k_1\times k_2}{\Sigma_{12}} = (\sigma_{ij}^{(12)}) \ , \ \sigma_{ij}^{(12)} = \{F(\xi_i,\eta_j)-p_iq_j\}/f_1(\xi_i)f_2(\eta_j),$$

$$i=1,\ldots,k_1 \ ; \ j=1,\ldots,k_2,$$

$$\underset{k_2\times k_1}{\Sigma_{21}} = \Sigma_{12}' \quad \text{and}$$

$$(5.71) \qquad \underset{k_2\times k_2}{\Sigma_{22}} = (\sigma_{ij}^{(22)}), \ \sigma_{ij}^{(22)} = q_i(1-q_j)/f_2(\eta_i)f_2(\eta_j) \quad \text{for } i \leqq j$$

$$= q_j(1-q_i)/f_2(\eta_i)f_2(\eta_j) \quad \text{for } i \geqq j \ .$$

From this theorem, we can have immediately an asymptotic joint distribution of $F(z_i^{(1)},z_j^{(2)})$'s.

Corollary of Theorem 5.2. Under the same notations and assumptions as in Theorem 5.2, the joint distribution of $w_{ij} = \sqrt{n}\ \{F(z_i^{(1)},z_j^{(2)}) - F(\xi_i,\eta_j)\}$, $i=1,\ldots,k_1$; $j=1,\ldots,k_2$, tends to the $k_1 \cdot k_2$ dimensional normal distribution with means 0 and covariances

$$
\begin{aligned}
\mathcal{E}(w_{ij}w_{k\ell}) &= a_{ik}^{(11)} p_{L_1}(1-p_{M_1})/f_1(\xi_i)f_1(\xi_k) + a_{i\ell}^{(12)}\{F(\xi_i,\eta_\ell)-p_i q_\ell\}/f_1(\xi_i)f_2(\eta_\ell) \\
&\quad + a_{jk}^{(12)}\{F(\xi_k,\eta_j)-p_k q_j\}/f_1(\xi_k)f_2(\eta_j) + a_{j\ell}^{(22)} q_{L_2}(1-q_{M_2})/f_2(\eta_j)f_2(\eta_\ell),
\end{aligned}
\tag{5.72}
$$

where

$$
a_{ik}^{(11)} = \left.\frac{\partial F}{\partial z_i^{(1)}}\frac{\partial F}{\partial z_k^{(1)}}\right|_{\substack{z^{(1)}=\xi \\ z^{(2)}=\eta}}, \quad a_{i\ell}^{(12)} = \left.\frac{\partial F}{\partial z_i^{(1)}}\frac{\partial F}{\partial z_\ell^{(2)}}\right|_{\substack{z^{(1)}=\xi \\ z^{(2)}=\eta}},
$$

$$
a_{jk}^{(12)} = \left.\frac{\partial F}{\partial z_k^{(1)}}\frac{\partial F}{\partial z_j^{(2)}}\right|_{\substack{z^{(1)}=\xi \\ z^{(2)}=\eta}},
$$

$$
a_{j\ell}^{(22)} = \left.\frac{\partial F}{\partial z_j^{(2)}}\frac{\partial F}{\partial z_\ell^{(2)}}\right|_{\substack{z^{(1)}=\xi \\ z^{(2)}=\eta}}, \quad \underline{\text{and}} \quad L_1 = \min(i,k),\ M_1 = \max(i,k),
$$

$L_2 = \min(j,\ell), M_2 = \max(j,\ell)$.

Remark: It is easy to conjecture that the similar argument as in Theorem 5.2 and its corollary would lead to the generalizations of our results for the two dimensional case to cases of more than two dimensions. For example, let us consider m dimensional case, and let $\xi_i^{(\alpha)}$ be the population quantile of order $p_i^{(\alpha)}$ with respect to the α^{th} component where $i=1,\dots,k_\alpha$; $\alpha = 1,\dots,m$ and let $z_i^{(\alpha)}$ be the corresponding quantity in samples of size n. Then, under the similar assumptions as in Theorem 5.2, the joint distribution of $z_i^{(\alpha)}$,

$i=1,\ldots,k_\alpha$; $\alpha = 1,\ldots,m$, would tend to a $k = \sum_{\alpha=1}^{m} k_\alpha$ dimensional normal distribution with means $\xi_i^{(\alpha)}$ and with the covariance matrix of the following form:

$$\frac{1}{n}\Sigma = \begin{bmatrix} \Lambda_{11} & \Lambda_{12} & \cdots & \Lambda_{1m} \\ \Lambda_{21} & \Lambda_{22} & \cdots & \Lambda_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ \Lambda_{m1} & \Lambda_{m2} & \cdots & \Lambda_{mm} \end{bmatrix},$$

where $\Lambda_{\alpha\alpha}$ is the $k_\alpha \times k_\alpha$ covariance matrix of $z_1^{(\alpha)},\ldots,z_{k_\alpha}^{(\alpha)}$ and has a form like (5.69) or (5.71) and $\Lambda_{\alpha\beta}$ $(\alpha \neq \beta)$ is the $k_\alpha \times k_\beta$ matrix having a form like (5.70).

After the finished draft, the author noticed Siddiqui's abstract [7] in which he treated the asymptotic distribution of a single quantile in the bivariate case.

REFERENCES

[1] Cramér, H. (1946). Mathematical Methods of Statistics. Princeton University Press.

[2] Mardia, K.V. (1963). "Correlation coefficient between ranges in samples from a bivariate population with applications to normal and Pareto type I populations, (abstract)." Ann. Math. Statist., **34**, 1131.

[3] Mardia, K.V. (1963). "Exact distribution of order statistics in samples from a multivariate population with applications to Pareto type I population, (abstract)." Ann. Math. Statist., **34**, 1131.

[4] Mood, A.M. (1941). "On the joint distribution of the median in samples from a multivariate population." Ann. Math. Statist., **12**, 268-278.

[5] Mosteller, F. (1946). "On some useful 'inefficient' statistics." Ann. Math. Statist., **17**, 377-407.

[6] Siddiqui, M.M. (1960). "Joint distribution of medians in samples from a bivariate population, (abstract)." Ann. Math. Statist., **31**, 528.

[7] Siddiqui, M.M. (1960), "Distributions of quantiles in samples from a bivariate population, (abstract)." Ann. Math. Statist., **31**, 820-821.

[8] Tippett, L.H.C. (1925). "On the extreme individuals and the range of samples taken from a normal population." Biometrika, **17**, 364-387.

A DECOMPOSITION OF DICHOTOMOUS RANDOM VARIABLES INTO INDEPENDENT COMPONENTS OF VARIATION

by

Juha Partanen

1. Introduction and Summary.

In the following we consider sets of variables which may assume only two different values representing, for example, the outcomes of some sociological or psychological measurements. The values are arbitrarily denoted by 0 and 1. These numbers are not interpreted, however, as real integers, but as elements of the field F_2 of two elements. Addition and multiplication are thus defined by the following table:

Addition

+	0	1
0	0	1
1	1	0

Multiplication

.	0	1
0	0	0
1	0	1

Correspondingly the sum of two dichotomous variables $\underline{x}$,[1] $\underline{y}$ will be defined by

$$\begin{aligned} \underline{x} + \underline{y} &= 1 , \quad \text{if } \underline{x} \neq \underline{y} \\ &= 0 , \quad \text{if } \underline{x} = \underline{y} , \end{aligned}$$

and whenever we consider sums of dichotomous variables later, they are defined in this way.

---

[1] In this paper, random variables are denoted by underlined small letters.

As a simple illustration, consider an observed variable x, which is composed of a "real" variable z and an "error" variable e. If both z and e are dichotomous, we may write $x = z + e$. For, if $e = 0$ then $x = z$, and if $e = 1$, $x \neq e$. The occurrence of error thus changes the value of the observed variable. More generally, we shall investigate the possibility of decomposing observed dichotomous variables into sums of latent dichotomous variables which represent various sources of variation.

Treating latent variables as dichotomous distinguishes this work from earlier investigations. In Lazarsfeld's latent structure analysis, for example, the underlying space is taken to be either a discrete and unordered set or a Euclidean space. By introducing latent dichotomous variables a certain symmetry is achieved which facilitites mathematical manipulations without changing conceptual considerations.

It may also be noted, that the approach used here bears some formal resemblance to the notion of symmetric parameters, as defined by Lazarsfeld and Bahadur, independently of each other, in (3). Here too, the joint distribution of dichotomous variables is described in terms of a symmetric set of parameters, although the basic mathematical formulation is quite different.

Summarizing the contents, Section 2 is devoted to basic algebra and Section 3 to basic results concerning the joint probability distribution of dichotomous variables. In Section 4 a measure of variation, the so-called v-measure is introduced. The decomposition into independent components is treated in Sections 5 and 6.

2. Algebraic concepts.

2.1. Boolean algebra.

Consider a set $B = \{x,y,z,\ldots\}$ which contains two special elements 0 and 1. Suppose that two operations on B are given, that is, two single-valued functions which assign to every ordered pair of elements of B an element of B. Assume that these operations, called addition and multiplication, satisfy the following requirements:

(i) B is an abelian group with respect to addition, having 0 as the identity element.

(ii) Multiplication is associative and commutative. The element 1 is a multiplicative unit and every element is idempotent with respect to multiplication, that is, $1x = x$ and $xx = x$ for every $x \in B$.

(iii) Multiplication is distributive with respect to addition.

Then B is called a Boolean algebra.

The basic example of Boolean algebra is provided by a collection of subsets of some set U, containing all unions, intersections, and complements of its sets. Such a collection contains in particular the empty set $\emptyset$ and the total set U. These sets take the role of the special elements 0 and 1, respectively, when symmetric set difference is taken for addition:

$$X + Y = (X \cap \overline{Y}) \cup (\overline{X} \cap Y) ,$$

and set intersection for multiplication:

$$XY = X \cap Y .$$

The validity of the requirements (i)-(iii) can be easily shown. Conversely, the original set operations can be expressed in terms of addition and multiplication, for we have

$$X \cup Y = X + Y + XY$$

$$\overline{X} = U + X.$$

It is convenient to use the notion of complement also in the abstract formulation. Thus we set $\bar{x} = 1+x$ for any $x \epsilon B$.

From the axioms (i)-(iii) a number of easily proven identities follow. Some of these are listed for further reference.

(1) $$x + x = 0$$

(2) $$x + y = x\bar{y} + \bar{x}y$$

(3) $$(\overline{x + y}) = xy + \bar{x}\bar{y}$$

(4) $$(\overline{xy}) = \bar{x}\bar{y} + \bar{x}y + x\bar{y}$$

The main reason for using symmetric set difference instead of set union as addition is that it trivially allows us to consider Boolean algebra as a linear space. For taking F_2, the field of two elements, as the coefficient field, and setting $0x = x$, $1x = x$, for every $x \epsilon B$, the axioms of linear space can be easily verified. This makes it possible to consider vectors with components from B. These vectors can be multiplied with matrices having 0 and 1 as elements. As for notations, the vectors will be treated as column vectors (in definitions they are usually written on a row and surrounded by parentheses).

Consider now a Boolean algebra B having a finite number of elements. A subset $\{x_1, \ldots, x_r\}$ of B is called a set of generators of B if any element of B can be expressed in terms of the elements

$x_1,\ldots,x_r$, using the operations addition, multiplication and complementation.

Given a set of generators $\{x_1,\ldots,x_r\}$ in a Boolean algebra B we define the basic elements as follows. An element of y is basic if it can be expressed as a product of r distinct factors, each factor being either x_i of $\overline{x}_i$. Clearly the number of distinct basic elements is 2^r.

Considering B as a linear space it can be shown that

Proposition 2.1. The non-zero basic elements, defined with respect to a set of generators $\{x_1,\ldots,x_r\}$ form a basis in B.

Proof. It must be shown that every element of B can be uniquely represented as a linear combination, that is, as a sum of non-zero basic elements; and that the non-zero basic elements are linearly independent.

It may be noted that the representation of an element as a sum of basic elements is known in propositional calculus as disjunctive normal form. Now, every element of B has an expression formed from sums and products of the generators $x_1,\ldots,x_r$ and from complements of these. Using the identities (3) and (4) one can decompose complemented sums and products. The parentheses can be removed by the distribution law. The result of these operations is a sum of products having either x_i's or $\overline{x}_i$'s as their factors. In each product, the missing x_i's can be added using the identity $x + \overline{x} = 1$. Applying the distribution law again, canceling duplicate terms, and deleting zero terms we finally get a unique representation as a sum of non-zero basic elements.

To show the linear independence of the non-zero basic elements $\{b_n\}$ assume contrarily that one of them can be represented as a sum of the others, say $b_m = \Sigma\, b_n$, $n \neq m$. Multiplying both sides by b_m we get the contradictory result $b_m = 0$, since the product of two distinct basic elements must always vanish.

2.2. Algebra of dichotomous variables.

Suppose a set of dichotomous variables $\underline{x}_1, \ldots, \underline{x}_r$ is given. They will be considered as two-valued functions defined on some probability field (the probability measure will be introduced later). Denoting the values of the variables arbitrarily by 0 and 1, we define addition of two variables by

$$\begin{aligned} \underline{x}_i + \underline{x}_j &= 1\,, \quad \text{if } \underline{x}_i \neq \underline{x}_j \\ &= 0\,, \quad \text{if } \underline{x}_i = \underline{x}_j \end{aligned}$$

and multiplication by

$$\begin{aligned} \underline{x}_i\underline{x}_j &= 1\,, \quad \text{if } \underline{x}_i = \underline{x}_j = 1 \\ &= 0\,, \quad \text{if } \underline{x}_i = 0 \text{ or } \underline{x}_j = 0\,. \end{aligned}$$

In particular we define the variables $\underline{1}$ and $\underline{0}$ to have the values 1 and 0, respectively, for every element of the probability field. The complement of a variable $\underline{x}$ will be denoted by $\bar{\underline{x}} = \underline{1} + \underline{x}$. It is easy to see that the axioms of Boolean algebra hold for the collection of the variables generated by $\underline{x}_1, \ldots, \underline{x}_r$. It will be denoted by $B(x_1, \ldots, x_r)$. It may be noted that if for each variable we consider those subsets of the probability field on which this variable has the value 1, we get an isomorphism between $B(x_1, \ldots, x_r)$ and a Boolean algebra of sets.

The basic elements of $B(x_1,\ldots,x_r)$ will be called *elementary variables*. They correspond to the points in the sample space of the variables $\underline{x}_1,\ldots,\underline{x}_r$.

In addition to the set of elementary variables there are other interesting sets, formed in a symmetrical way with respect to the generators $\underline{x}_1,\ldots,\underline{x}_r$. In this context we consider only the set of *sum variables* of the form $\underline{x}_{i_1}+\cdots+\underline{x}_{i_n}$. To exhibit the relationships between the elementary variables and the sum variables we use the notion of the Kronecker product.

Let $\underline{y}^m = (\underline{y}_1,\ldots,\underline{y}_m)$ and $\underline{z}^n = (\underline{z}_1,\ldots,\underline{z}_n)$ be two vectors with components in $B(x_1,\ldots,x_r)$. The *Kronecker product* of the vectors $\underline{y}^m$ and $\underline{z}^n$, denoted by $\underline{y}^m \times \underline{z}^n$, is defined as the mn-dimensional vector

$$\underline{y}^m \times \underline{z}^n = (\underline{y}_1\underline{z}_1,\underline{y}_2\underline{z}_1,\ldots,\underline{y}_m\underline{z}_1,\underline{y}_1\underline{z}_2,\ldots,\underline{y}_m\underline{z}_m).$$

It can be verified that this product is associative.

Lemma 2.1. Let $\underline{y}' = C\underline{y}$ and $\underline{z}' = D\underline{z}$, where $\underline{y}$, $\underline{z}$, $\underline{y}'$, $\underline{z}'$ are vectors and C, D matrices with appropriate dimensions. Then there exists a matrix, denoted by $C \times D$, such that $\underline{y}' \times \underline{z}' = (C \times D)(\underline{y} \times \underline{z})$.

To prove the lemma, note that a component of the vector $\underline{y}' \times \underline{z}'$ can be represented in the form

$$(\Sigma\, c_{mn}\underline{y}_n)(\Sigma\, d_{hk}\underline{z}_k) = \Sigma\Sigma\, c_{mn}d_{hk}\,\underline{y}_n\underline{z}_k)\ ,$$

that is, as a linear combination of the products $\underline{y}_n\underline{z}_k$, which are the components of $\underline{y} \times \underline{z}$. The matrix $C \times D$ is known as the Kronecker product of C and D. In compound matrix notation $C \times D$ can be written

$$C \times D = \begin{bmatrix} d_{11}C & d_{12}C & \cdots \\ d_{21}C & d_{22}C & \cdots \\ \vdots & \vdots & \end{bmatrix} .$$

Using the notion of Kronecker product we generate the elementary variables as follows. Let

$$\underline{b}^1 = \begin{bmatrix} \underline{\bar{x}}_1 \\ \underline{x}_1 \end{bmatrix}$$

and define $\underline{b}^i$ recursively by

$$\underline{b}^i = \underline{b}^{i-1} \times \begin{bmatrix} \underline{\bar{x}}_i \\ \underline{x}_i \end{bmatrix}, \ i=2,\ldots,r.$$

Clearly $\underline{b}^r$ is a 2^r-dimensional vector having the elementary variables of $B(x_1,\ldots,x_r)$ as components. Thus we have, for example,

$$b^2 = (\underline{\bar{x}}_1\underline{\bar{x}}_2,\ \underline{x}_1\underline{\bar{x}}_2,\ \underline{\bar{x}}_1\underline{x}_2,\ \underline{x}_1\underline{x}_2) \ .$$

The components of $\underline{b}^r$ will be denoted by $\underline{b}^r_m$, $m=0,1,\ldots,2^r-1$. Clearly $\underline{b}^r_m\,\underline{b}^r_n = \underline{0}$ whenever $m \neq n$ and $\Sigma\,\underline{b}^r_m = \underline{1}$.

The sum variables can be generated by a similar construction. Let

$$\underline{a}^1 = \begin{bmatrix} \underline{0} \\ \underline{x}_1 \end{bmatrix} ,\ \underline{\bar{a}}^1 = \begin{bmatrix} \underline{1} \\ \underline{\bar{x}}_1 \end{bmatrix}$$

Define the vectors $\underline{a}^i$, $i=2,\ldots,r$, recursively by

$$\underline{a}^i = \underline{a}^{i-1} \times \begin{bmatrix} \underline{1} \\ \underline{\bar{x}}_i \end{bmatrix} + \underline{\bar{a}}^{i-1} \times \begin{bmatrix} \underline{0} \\ \underline{x}_i \end{bmatrix}$$

$$\underline{\bar{a}}^i = \underline{\bar{a}}^{i-1} \times \begin{bmatrix} \underline{1} \\ \underline{\bar{x}}_i \end{bmatrix} + \underline{a}^{i-1} \times \begin{bmatrix} \underline{0} \\ \underline{x}_i \end{bmatrix} .$$

Clearly the vectors $\underline{a}^r$ and $\overline{\underline{a}}^r$ are 2^r-dimensional. Denote their components by $\underline{a}_m^r$ and $\overline{\underline{a}}_m^r$, $m=0,1,\ldots,2^r-1$, respectively. Using the identities (2) and (3) it can be verified and proven by induction that the components $\underline{a}_m^r$ are sum variables and $\overline{\underline{a}}_m^r$ their complements. In the case $r=2$ we have $\underline{a}^2 = (\underline{0}, \underline{x}_1, \underline{x}_2, \underline{x}_1 + \underline{x}_2)$.

It follows from proposition 2.1 that the sum variables as well as their complements can be represented as linear combinations of elementary variables. We investigate now this representation, denoting in general

$$\begin{aligned} \underline{a}^r &= A_r\underline{b}^r\ , \quad A_r = [a_{mn}^r] \\ \overline{\underline{a}}^r &= \overline{A}_r\underline{b}^r\ , \quad \overline{A}_r = [\overline{a}_{mn}^r]\ . \end{aligned} \tag{5}$$

To justify the notation we must show that $a_{mn}^r + \overline{a}_{mn}^r = 1$ for every m,r,r. Now a component $\overline{\underline{a}}_m^r = \underline{1} + \underline{a}_m^r = \Sigma\, \underline{b}_n^r + \Sigma\, a_{mn}^r\underline{b}_n^r = \Sigma(1+a_{mn}^r)\underline{b}_n^r$; therefore $\overline{a}_{mn}^r = 1+a_{mn}^r$.

When $r = 1$ we have

$$\underline{a}^1 = \begin{bmatrix} 0 \\ \underline{x}_1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} \overline{x}_1 \\ x_1 \end{bmatrix}.$$

Thus

$$A_1 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}\ ,\quad \overline{A}_1 = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}.$$

For A_r we get the recursion formula (using lemma 2.1):

$$\underline{a}^{r} = \underline{a}^{r-1} \times \begin{bmatrix} \underline{1} \\ \underline{x}_r \end{bmatrix} + \overline{\underline{a}}^{r-1} \times \begin{bmatrix} \underline{0} \\ \underline{x}_r \end{bmatrix}$$

$$= A_{r-1}\underline{b}^{r-1} \times \overline{A}_1 \begin{bmatrix} \overline{\underline{x}}_r \\ \underline{x}_r \end{bmatrix} + \overline{A}_{r-1}\underline{b}^{r-1} \times A_1 \begin{bmatrix} \overline{\underline{x}}_r \\ \underline{x}_r \end{bmatrix}$$

$$= (A_{r-1} \times \overline{A}_1 + \overline{A}_{r-1} \times A_1)\underline{b}^r .$$

Because $\underline{a}^r = A_r \underline{b}^r$ we have thus

$$A_r = A_{r-1} \times \overline{A}_1 + \overline{A}_{r-1} \times A_1 \, . \tag{6}$$

In compound matrix notation this can be written as

$$\begin{aligned} A_r &= \begin{bmatrix} A_{r-1} & A_{r-1} \\ A_{r-1} & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & \overline{A}_{r-1} \end{bmatrix} \\ &= \begin{bmatrix} A_{r-1} & A_{r-1} \\ A_{r-1} & \overline{A}_{r-1} \end{bmatrix} . \end{aligned} \tag{6'}$$

Correspondingly we get for $\overline{A}_r$ the recursion formula

$$\overline{A}_r = A_{r-1} \times A_1 + \overline{A}_{r-1} \times \overline{A}_1 \, , \tag{7}$$

or

$$\begin{aligned} \overline{A}_r &= \begin{bmatrix} 0 & 0 \\ 0 & A_{r-1} \end{bmatrix} + \begin{bmatrix} \overline{A}_{r-1} & \overline{A}_{r-1} \\ \overline{A}_{r-1} & 0 \end{bmatrix} \\ &= \begin{bmatrix} \overline{A}_{r-1} & \overline{A}_{r-1} \\ \overline{A}_{r-1} & A_{r-1} \end{bmatrix} . \end{aligned} \tag{7'}$$

Similar results could be obtained for the product variables of the form $\underline{1}$, $\underline{x}_1$, $\underline{x}_2$, $\underline{x}_1\underline{x}_2, \ldots$. They will not be needed, however, in what follows.

3. The joint distribution of dichotomous variables.

We now introduce the probability measure into our considerations. Let the dichotomous variables $\underline{x}_1,\ldots,\underline{x}_r$ be given. The joint probability distribution is usually specified by giving the probabilities of the elementary events. Because their sum always equals one, 2^r-1 real-valued parameters are needed for the complete specification of the joint distribution.

In the following, the probability of a dichotomous variable means the probability that the variable has the value 1. We denote the vector of the probabilities of the elementary variables by p^r,

$$p^r = (p_o^r,\ p_1^r,\ldots,p_{2^r-1}^r)$$

$$p_m^r = P(\underline{b}_m^r = 1);\ 0 \leq p_m^r \leq 1\ ;\ \sum_m p_m^r = 1\ ;$$

and the vector of the probabilities of the sum variables by q^r,

$$q^r = (q_o^r,\ q_1^r,\ldots,q_{2^r-1}^r)$$

$$q_m^r = P(\underline{a}_m^r = 1)\ ;\ 0 \leq q_m^r \leq 1\ .$$

The probability vector of the complements of the sum variables is denoted by $\bar{q}^r = (\bar{q}_o^r,\ldots,\bar{q}_{2^r-1}^r)$.

According to the definition, probability measure is additive with respect to non-intersecting sets. It follows that $P(\underline{x} + \underline{y} = 1) = P(\underline{x} = 1) + P(\underline{y} = 1)$ whenever $\underline{x}\ \underline{y} = \underline{0}$. This allows us to derive a representation for q^r and $\bar{q}^r$ as a function of p^r. For according to the representation (2.5) any sum variable is a sum of elementary variables. Because $\underline{b}_m^r\,\underline{b}_n^r = \underline{0}$ whenever $m \neq n$, the probability of a sum variable is the corresponding sum of probabilities of elementary variables. Defining the matrices with real-valued elements 0 and 1 in the same places as in the matrices A_r and $\bar{A}_r$, and using the same notations for these (this should not cause confusion) we have the following proposition:

Proposition 3.1. The probability vector q^r and its complement $\bar{q}^r$ of the sum variables have the representation

$$q^r = A_r p^r \ ; \ \bar{q}^r = \bar{A}_r p^r$$

as a function of the probability vector of the elementary variables.

The inverse for this representation will be derived in 3.2.

Consider the case of two variables more closely. The joint distribution is often given as a four-fold table,

$x_1 \backslash x_2$	0	1	
0	p_{00}	p_{01}	$\bar{q}_1$
1	p_{10}	p_{11}	q_1
	$\bar{q}_2$	q_2	1

so that

$$\begin{aligned} q_1 &= p_{10} + p_{11} \\ q_2 &= p_{01} + p_{11} \\ q_{12} &= p_{01} + p_{10} \end{aligned} \tag{1}$$

For this particular case we adopt a simpler index system, denoting $q_1 = q^2_{10}$, $q_2 = q^2_{01}$, $q_{12} = q^2_{11}$.

We now prove

Proposition 3.2. Let q_1, q_2, q_{12} be real numbers between 0 and 1. The conditions

$$q_{12} \geq |q_1 - q_2|$$

$$\bar{q}_{12} \geq |q_1 - \bar{q}_2| \; ; \; \bar{q}_{12} = 1 - q_{12}, \; \bar{q}_2 = 1 - q_2$$

are necessary and sufficient for q_1, q_2 to be probabilities of two dichotomous variables, and q_{12} the probability of their sum.

Proof. To show the necessity we note that $|q_1-q_2| = |p_{01}-p_{10}| \leq p_{01}+ p_{10} = q_{12}$ and $|q_1-\bar{q}_2| = |p_{11}-q_{00}| \leq p_{11} + p_{00} = \bar{q}_{12}$ according to the equations (1).

To prove the converse we construct a four-fold table for which the equations (1) hold. Suppose e.g. $q_1 \geq q_2$. Now $q_{12} = p_{01} + p_{10} = p_{01} + (q_1-q_2+ p_{01}) = q_1-q_2+ 2p_{01}$. Accordingly define

$$p_{01} = \frac{1}{2}(-q_1 + q_2 + q_{12}).$$

On the basis of (1) define also

$$p_{10} = \frac{1}{2}(q_1-q_2+ q_{12})$$
$$p_{00} = \frac{1}{2}(-q_1+ \bar{q}_2+ \bar{q}_{12})$$
$$p_{11} = \frac{1}{2}(q_1-\bar{q}_2+ \bar{q}_{12}) \; .$$

The conditions of the proposition ensure that the elementary probabilities thus defined are ≥ 0. Also the equations (1) hold, and it is readily checked that their sum equals one.

The inequalities of the proposition 3.2 hold whenever one of the variables is the sum of the two others. Thus we also have

$$\begin{cases} q_1 \geq |q_2 - q_{12}| , \\ \bar{q}_1 \geq |q_2 - \bar{q}_{12}| , \end{cases} \qquad \begin{cases} q_2 \geq |q_1 - q_{12}| \\ q_2 \geq |q_1 - q_{12}| . \end{cases}$$

The geometric equivalent to each of the three pairs of inequalities is the interior of the tetrahedron with vertices at the points (0,0,0), (0,1,1), (1,0,1), (1,1,0) of the (q_1, q_2, q_{12})-space.

3.2 Hadamard matrices.

Define the matrices H_r, $r=1,2,\ldots,$ with elements +1 and -1 (real numbers) as

$$H_r = \bar{A}_r - A_r , \; r=1,2,\ldots .$$

The matrix H_r is $(2^r \times 2^r)$-dimensional. In particular we have

$$H_1 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} ,$$

$$H_2 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$$

In general we have the recursion formula

$$\begin{aligned} H_{r+1} &= \bar{A}_{r+1} - A_{r+1} \\ &= \begin{bmatrix} \bar{A}_r - A_r & \bar{A}_r - A_r \\ \bar{A}_r - A_r & A_r - \bar{A}_r \end{bmatrix} \\ &= \begin{bmatrix} H_r & H_r \\ H_r & -H_r \end{bmatrix} \\ &= H_r \times H_1 . \end{aligned}$$

These matrices, called Hadamard matrices, are used e.g., in coding theory (cf [2], p. 79) and in planning of experiments. We need some of their properties later.

Proposition 3.3. H_r is symmetric and $H_r^{-1} = 2^{-r}H_r$.

Proof. Symmetry follows immediately from the fact that H_1 is symmetric and that this property is preserved in going from r to $r+1$. To prove the formula for the inverse we have

$$H_1^2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

$$= 2\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

and

$$H_{r+1}^2 = \begin{bmatrix} H_r & H_r \\ H_r & -H_r \end{bmatrix}\begin{bmatrix} H_r & H_r \\ H_r & -H_r \end{bmatrix}$$

$$= 2\begin{bmatrix} H_r^2 & 0 \\ 0 & H_r^2 \end{bmatrix} .$$

Thus, if $H_r^2 = 2^rI$, the same holds for $r+1$; so that in general $H_r^2 = 2^rI$. The assertion follows from this.

From the symmetry and the relation $H_rH_r = 2^rI$ it follows that H_r, normalized by $2^{-r/2}$ is an orthogonal matrix. It may be also noted that its first row consists of 1's only.

From the proposition 3.1 it follows that

$$\bar{q}^r - q^r = \bar{A}_rp^r - A_rp^r$$

$$= H_rp^r .$$

6

Applying the proposition 3.3 we get

Corollary 3.3.1. The probability vector of the elementary variables can be expressed as

$$p^r = 2^{-r} H_r(\bar{q}^r - q^r) .$$

This result, together with the proposition 3.1, shows that there is a one-one correspondence between the probabilities of sum variables and those of elementary variables.

4. A measure of variation.

We are now going to define a measure for dichotomous variables which has the property of being additive with respect to (finite) sums of independent variables.. In comparison it may be noted that probability measure is additive with respect to sums of non-intersecting variables.

Suppose $\underline{x}_1$ and $\underline{x}_2$ are independent dichotomous variables. The probability of $\underline{x}_1 + \underline{x}_2$ is then

$$\begin{aligned} q_{12} &= P\{\underline{x}_1 + \underline{x}_2 = 1\} \\ &= P\{\underline{x}_1 \bar{\underline{x}}_2 + \bar{\underline{x}}_1 \underline{x}_2 = 1\} \\ &= P\{\underline{x}_1 \bar{\underline{x}}_2 = 1\} + P\{\bar{\underline{x}}_1 \underline{x}_2 = 1\} \\ &= q_1\bar{q}_2 + \bar{q}_1 q_2 \ , \end{aligned}$$

where $q_i = P\{\underline{x}_i = 1\}$, $\bar{q}_i = P\{\bar{\underline{x}}_i = 1\}$, $i=1,2$. Similarly

$$\bar{q}_{12} = q_1 q_2 + \bar{q}_1\bar{q}_2 \ .$$

It follows that

$$q_{12}-\bar{q}_{12} = -(q_1-\bar{q}_1)(q_2-\bar{q}_2) \ . \tag{1}$$

We now give the following definition.

Definition. The v-measure of a dichotomous random variable $\underline{x}$ is given by the expression

$$v(x) = -\log |p-\bar{p}| \ , \ p = P\{\underline{x} = 1\} \ . \tag{2}$$

For the base of logarithm, any fixed real number larger than one can be chosen.

Proposition 4.1. The v-measure has the following properties:

i) $v(x) \geq 0$;

ii) $v(x) = v(\bar{x})$;

iii) $v(x)$ increases monotonically from zero to $+\infty$ when p increases from zero to $\frac{1}{2}$ or decreases from one to $\frac{1}{2}$.

Moreover, if $\underline{x}_1$, $\underline{x}_2$ are independent,

$$v(x_1 + x_2) = v(x_1) + v(x_2) . \tag{3}$$

Proof. The three properties follow immediately from the definition. To prove (3), take the logarithm of the absolute value on both sides of (1).

It can be shown that the additivity condition (3) makes the v-measure essentially unique. For using probabilities as arguments, the formula (3) can be written

$$v(q_1\bar{q}_2 + \bar{q}_1q_2) = v(q_1) + v(q_2) .$$

Supposing that v is differentiable we get $v'(q_1\bar{q}_2 + \bar{q}_1q_2)(\bar{q}_1 - q_1) = v'(q_2)$. Setting $q_2 = 0$ we have $v'(q_1)(\bar{q}_1 - q_1) = v'(0)$, and the solution of this is of the form (2).

Thus the v-measure operates in a manner which measures the difficulty which is attached to the guessing of the right value of $\underline{x}$. When the probability of $\underline{x}$ is one or zero, $v(x)$ is zero; when the probability is $1/2$, $v(x)$ is infinite or we may say guessing is "infinitely difficult".

The discontinuity of v-measure at $p = 1/2$ is a disadvantage but may not really affect its use in applications. In the following we assume that the probabilities of all the variables under consideration are $\neq \frac{1}{2}$.

We now define the variation vector or v-vector α^r of the given dichotomous variables $\underline{x}_1, \ldots, \underline{x}_r$ as the 2^r-dimensional vector of v-measures of the sum variables $\underline{a}^r_m$, $m=0,\ldots,2^r-1$. Thus

$$\alpha^r = (\alpha^r_o, \ldots, \alpha^r_{2^r-1})$$

$$\alpha^r_m = v(a^r_m)$$

$$= -\log |q^r_m - \bar{q}^r_m| .$$

The variation vector α^r is non-singular if all its components are $\neq +\infty$.

Consider the case $r = 2$ in detail. Using the same notations as on page 12 we have

$$\alpha^2_o = 0$$

$$\alpha^2_1 = -\log |q_1 - \bar{q}_1|$$

$$\alpha^2_2 = -\log |q_2 - \bar{q}_2|$$

$$\alpha^2_{12} = -\log |q_{12} - \bar{q}_{12}| .$$

If the variables are completely dependent on each other we have the following four-fold table and v-vector:

$$\begin{array}{c|cc|c} x_1 \;\; {}^{x_2} & 0 & 1 & \\ \hline 0 & \bar{q} & 0 & \bar{q} \\ 1 & 0 & q & q \\ \hline & \bar{q} & q & 1 \end{array} \; ; \quad \alpha^2 = \begin{bmatrix} 0 \\ \alpha \\ \alpha \\ 0 \end{bmatrix} \; ; \quad \alpha = -\log |q - \bar{q}| .$$

In case the variables are independent we get

x_1 \ x_2	0	1	
0	$\bar{q}_1\bar{q}_2$	$\bar{q}_1 q_2$	$\bar{q}_1$
1	$q_1\bar{q}_2$	$q_1 q_2$	q_1
	$\bar{q}_2$	q_2	1

$$; \alpha^2 = \begin{bmatrix} 0 \\ \alpha_1 \\ \alpha_2 \\ \alpha_1 + \alpha_2 \end{bmatrix} ; \alpha_i = -\log |q_i - \bar{q}_i| ; i=1,2.$$

It may be asked whether the variation vector determines the distribution uniquely. This happens only in special cases, in general there is more than one distribution corresponding to a given variation vector. This is due to the fact that the variable and its complement have the same v-measure. Thus for a particular value of v-measure usually one can choose either one of the two possible complementary probabilities.

According to this, the number of distributions corresponding to a given v-vector would be $(2)^{2^r-1}$. However, all these may not be possible. For it follows from proposition 3.2 that whenever $|q_1-q_2| > \frac{1}{2}$ or $|q_1-\bar{q}_2| > \frac{1}{2}$ then $q_{12} > \frac{1}{2}$ or $q_{12} < \frac{1}{2}$ respectively. That is, the choice of q_1 and q_2 determines q_{12} uniquely, and thus the number of distributions corresponding to a given v-vector is generally $\leq (2)^{2^r-1}$.

As an illustration consider the following numerical example.

x_1 \ x_2	0	1	
0	.15	.15	.30
1	.25	.45	.70
	.40	.60	

x_1 \ x_2	0	1	
0	.05	.25	.30
1	.35	.35	.70
	.40	.60	

In both tables we have $q_1 = .70$, $q_2 = .60$ and $|q_{12}-\bar{q}_{12}| = .20$. Permuting rows and columns we get eight distributions which all have the same v-vector. The distributions above look quite dissimilar, however, and the conventional measures of association assume different values. On the other hand, choosing e.g., $q_1 = .75$, $q_2 = .80$ we would have only four different distributions because in this case necessarily $\bar{q}_{12} \geq |q_1-\bar{q}_2| = .55$.

5. Decomposition into covariables.

Consider first two dichotomous variables $\underline{x}_1$ and $\underline{x}_2$ with an arbitrary joint distribution. In the following we shall be concerned with the decomposition of the form

(1)
$$\underline{x}_1 = \underline{z}_1 + \underline{z}_{12} ,$$
$$\underline{x}_2 = \underline{z}_2 + \underline{z}_{12} ,$$

where $\underline{z}_1$, $\underline{z}_2$, $\underline{z}_{12}$ are dichotomous variables. The variable $\underline{z}_{12}$ is interpreted as the "common part" or the covariable of $\underline{x}_1$ and $\underline{x}_2$, the terms $\underline{z}_1$ and $\underline{z}_2$ represent the "unique variation" of $\underline{x}_1$ and $\underline{x}_2$ respectively. From (1) it follows immediately that $\underline{x}_1 + \underline{x}_2 = \underline{z}_1 + \underline{z}_2$ so that $\underline{z}_i$ can be interpreted as the covariable of $\underline{x}_i$ and $\underline{x}_1 + \underline{x}_2$, i=1,2.

The corresponding decomposition for the three variables $\underline{x}_1$, $\underline{x}_2$, $\underline{x}_3$ has the form

(2)
$$\underline{x}_1 = \underline{z}_1 + \underline{z}_{12} + \underline{z}_{13} + \underline{z}_{123}$$
$$\underline{x}_2 = \underline{z}_2 + \underline{z}_{12} + \underline{z}_{23} + \underline{z}_{123}$$
$$\underline{x}_3 = \underline{z}_3 + \underline{z}_{13} + \underline{z}_{23} + \underline{z}_{123}$$

For each non-empty set of variables there is thus a corresponding covariable. Each observed variable is represented as the sum of the covariables corresponding to the sets which include the variable in question.

The equations (1) and (2) are purely formal as long as no assumptions about the joint distribution of the covariables has been made. In this case the assumption of independence is the most natural. We thus require that the covariables are totally independent of each other. It follows immediately that the v-measure of $\underline{x}_1$ can be represented as a sum of the v-measures of the covariables. The equations (1) and (2) thus imply corresponding equations for the v-measures.

It is by no means certain that the decompositions (1) and (2) are generally valid when the independence assumption is made. In fact we shall see that they are not, and the approach will be modified accordingly. But now we proceed on the basis of (1) and (2).

The equations (1) and the derived expression for $\underline{x}_1 + \underline{x}_2$ can be written

$$\begin{aligned} \underline{0} &= \underline{0} \\ \underline{x}_1 &= \underline{z}_1 + \underline{z}_{12} \\ \underline{x}_2 &= \underline{z}_2 + \underline{z}_{12} \\ \underline{x}_1 + \underline{x}_2 &= \underline{z}_1 + \underline{z}_2 \, , \end{aligned}$$

or, using vector and matrix notation,

$$\underline{a}^2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \underline{0} \\ \underline{z}_1 \\ \underline{z}_2 \\ \underline{z}_{12} \end{bmatrix}$$

$$= \quad A_2\underline{z}^2 \quad , \quad \underline{z}^2 = (\underline{0}, \underline{z}_1, \underline{z}_2, \underline{z}_{12}) .$$

Similarly it can be shown that the equations (3) and those derived from them are represented by

$$\underline{a}^3 = A_3\underline{z}^3 \, , \, \underline{z}^3 = (\underline{0}, \underline{z}_1, \underline{z}_2, \underline{z}_{12}, \underline{z}_3, \underline{z}_{13}, \underline{z}_{23}, \underline{z}_{123})$$

Generally we thus define the vector $\underline{z}^r$ of the covariables by the equation

(3) $$\underline{a}^r = A_r\underline{z}^r \ , \ \underline{z}^r = (\underline{z}_0^r, \ldots, \underline{z}_{2^r-1}^r) \ .$$

It may be observed that the equation (3) is formally similar to (2.5) which defined the relationship between the sum variables and the elementary variables. The basic difference is that the covariables are assumed independent whereas the elementary variables were non-intersecting.

5.1. Components of covariation.

Consider now the v-measures of the variables defined above. Let α^r be the variation vector of the variables $\underline{x}_1, \ldots, \underline{x}_r$. We assume that α^r is non-singular. The vector of the v-measures of the covariables is denoted by γ^r,

$$\gamma^r = (\gamma_o^r, \ldots, \gamma_{2^r-1}^r)$$

$$\gamma_o^r = 0$$

$$\gamma_m^r = -\mathrm{Log}|P(\underline{z}_m^r = 1) - P(\underline{z}_m^r = 0)|, \ m=1,\ldots,2^r-1 \ .$$

The components γ_m^r are called components of covariation.

From the independence of the covariables it follows by the theorem 4.1, that the v-measure of a sum of covariables equals the sum of the corresponding components of covariation. On the basis of (3) we thus get

(4) $$\alpha^r = A_r\gamma^r \ ,$$

which defines the relationship between α^r and γ^r. The matrix A_r is now interpreted as a matrix with real-valued elements.

To solve for the γ^r we use the following lemma.

Lemma 5.1. The product H_rA_r can be written:

$$H_rA_r = 2^{r-1}\begin{bmatrix} 0 & 1 & 1 & .. & 1 \\ 0 & -1 & 0 & .. & 0 \\ 0 & 0 & -1 & .. & 0 \\ . & . & . & . & . \\ 0 & 0 & 0 & .. & -1 \end{bmatrix}$$

Proof. Form

$$H_1A_1 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 1 \\ 0 & -1 \end{bmatrix}$$

and

$$H_{r+1}A_{r+1} = \begin{bmatrix} H_r & H_r \\ H_r & -H_r \end{bmatrix}\begin{bmatrix} A_r & A_r \\ A_r & \bar{A}_r \end{bmatrix}$$

$$= \begin{bmatrix} 2H_rA_r & H_r[1] \\ 0 & H_r(A_r-\bar{A}_r) \end{bmatrix},$$

where by $[1] = A_r + \bar{A}_r$ we denote the matrix with 1 in each position. It can be verified that $H_r[1]$ is a matrix with 2^r in each position in the first row and zeros elsewhere (this follows from the orthogonality of H_r, whose first row consists of 1's only). Also $H_r(A_r-\bar{A}_r) = -H_r^2 = -2^rI$ by the proposition 3.2. Thus, if the assertion holds for r it also holds for r+1.

Multiplying both sides of (4) by H_r and applying the lemma 5.1 we get

(5) $$H_r\alpha^r = H_rA_r\gamma^r \,, \quad H_r = [h_{mn}]$$

(6) $$\gamma^r_m = -2^{-(r-1)} \sum_n h_{mn}\, \alpha^r_n \,, \quad m=1,\dots,2^r-1$$

and for the first row

(7) $$\sum \gamma^r_m = 2^{-(r-1)} \sum \alpha^r_r \,.$$

Let

$$T = 2^{-(r-1)} \sum \alpha^r_m$$

and call T the measure of total variation. Thus

(7)' $$T = \sum \gamma^r_m \,.$$

Defining the vector $\gamma^r(T)$ by

$$\gamma^r(T) = (-T, \gamma^r_1, \dots, \gamma^r_{2^r-1})$$

we get an alternative form for the equation (5):

(5)' $$\gamma^r(T) = -\frac{1}{2^{r-1}} H_r\alpha^r \,.$$

As an illustration, consider the case $r = 2$. The equation (5)' becomes

$$\begin{bmatrix} -T \\ \gamma_1 \\ \gamma_2 \\ \gamma_{12} \end{bmatrix} = -\frac{1}{2} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ \alpha_1 \\ \alpha_2 \\ \alpha_{12} \end{bmatrix} ,$$

so that

$$T = \frac{1}{2}(\alpha_1 + \alpha_2 + \alpha_{12})$$

$$\gamma_1 = \frac{1}{2}(\alpha_1 - \alpha_2 + \alpha_{12})$$

$$\gamma_2 = \frac{1}{2}(-\alpha_1 + \alpha_2 + \alpha_{12}) \tag{8}$$

$$\gamma_{12} = \frac{1}{2}(\alpha_1 + \alpha_2 - \alpha_{12}) .$$

In particular, if $\underline{x}_1$ and $\underline{x}_2$ are independent, $\alpha_{12} = \alpha_1 + \alpha_2$. Then $\gamma_{12} = 0$ and $\gamma_1 = \alpha_1$, $\gamma_2 = \alpha_2$. On the other hand, if $\underline{x}_1$ and $\underline{x}_2$ are completely dependent on each other, $\alpha_1 = \alpha_2$ and $\alpha_{12} = 0$, so that we have $\gamma_{12} = \alpha_1 = \alpha_2$ and $\gamma_1 = \gamma_2 = 0$. Thus the whole variation is placed on the covariation component of $\underline{x}_1$ and $\underline{x}_2$, whereas in the first case only the components of unique variation are non-zero. It will be seen that these results hold for general r.

5.2. Two special cases.

Until now, the introduction of covariables and the corresponding components of covariation has hardly been motivated at all, except by noting that the algebraic relationship between the sum variables and the covariables is identical with the relationship between sum variables and elementary variables. By considering two special cases we wish to show now that the components of covariation, defined whenever the variation vector of $\underline{x}_1,\dots,\underline{x}_r$ is non-singular, describe the joint distribution of $\underline{x}_1,\dots,\underline{x}_r$ in a meaningful way.

Proposition 5.1. Let $\underline{x}_1,\dots,\underline{x}_r$ be completely dependent on each other so that $P\{\underline{x}_i = \underline{x}_j\} = 0$ or 1 for all i,j, and denote $\alpha = v(x_i)$. We then have

$$T = \alpha$$

$$\gamma^r_m = 0 \quad , \quad m < 2^r - 1$$

$$\gamma^r_{2^r-1} = \alpha \quad ,$$

so that all components of covariation vanish, except the one corresponding to the covariation of all $\underline{x}_1, \ldots, \underline{x}_r$.

Proof. The result has been proved for $r=2$. In the general case it is easy to see that a component $\alpha^r_m = \alpha$ where m is the index of a sum variable having odd number of terms, and $\alpha^r_m = 0$ otherwise. Consider now α^{r+1}. Represent it in the form,

$$\alpha^{r+1} = \begin{bmatrix} \alpha^r \\ \alpha[1] - \alpha^r \end{bmatrix} ,$$

where $[1]$ is the 2^r-dimensional vector of 1's in each position. The validity of this representation follows from the fact that adding $\underline{x}_{r+1}$ to each $\underline{a}^r_m$ increases the number of terms by one and thus changes α to 0, and conversely. Applying (5)' we get

$$\gamma^{r+1}(T) = -\frac{1}{2^r} \begin{bmatrix} H_r & H_r \\ H_r & -H_r \end{bmatrix} \begin{bmatrix} \alpha^r \\ \alpha[1] - \alpha^r \end{bmatrix}$$

$$= -\frac{1}{2^r} \begin{bmatrix} \alpha H_r[1] \\ 2H_r\alpha^r - \alpha H_r[1] \end{bmatrix} .$$

But $H_r[1]$ is the vector with 2^r in the first row and zeros elsewhere. Assuming that the desired result holds for r, we can see that all the components of $\gamma^{r+1}(T)$ vanish, except the first and the last, which both equal α. Hence the result holds generally.

Proposition 5.2. Let $\underline{x}_1,\dots,\underline{x}_r$ be totally independent of each other, and denote $\alpha_i = v(x_i)$, $i=1,\dots,r$. Then $T = \sum \alpha_i$ and all the components of covariation vanish, except those corresponding to unique variation of each $\underline{x}_i$. Their values are equal to the v-measures of the corresponding $\underline{x}_i$.

Proof. The assertion holds when $r=2$. Consider now α^{r+1}. Because $\underline{x}_{r+1}$ is totally independent of the $\underline{x}_1,\dots,\underline{x}_r$, it is also independent of their sums. From the additivity of the v-measure with respect to independent variables it follows that α^{r+1} can be represented in the form

$$\alpha^{r+1} = \begin{bmatrix} \alpha^r \\ \alpha^r + \alpha_{r+1}[1] \end{bmatrix},$$

where [1] denotes the 2^r-dimensional vector of 1's. Applying (5)' we get

$$\gamma^{r+1}(T) = -\frac{1}{2^r}\begin{bmatrix} H_r & H_r \\ H_r & -H_r \end{bmatrix}\begin{bmatrix} \alpha^r \\ \alpha^r + \alpha_{r+1}[1] \end{bmatrix}$$

$$= -\frac{1}{2^r}\begin{bmatrix} 2H_r\alpha^r + \alpha_{r+1}H_r[1] \\ \alpha_{r+1}H_r[1] \end{bmatrix}$$

It can now be verified that if the assertion holds for r, then it also holds for $r+1$. In particular we have $\gamma^{r+1}_{2^r} = \alpha_{r+1}$, $\gamma^{r+1}_{2^r}$ being the component of the unique variation of $\underline{x}_{r+1}$.

5.3. A numerical example.

The data for this example comes from the paper [1] p. 340. The variables are four dichotomized tests measuring engineering ability. The number of subjects is 137.

The discussion above was carried through in terms of theoretical probabilities. In most practical situations only sample estimates of these are known. It would thus be necessary to consider procedures for the estimation of the components of covariation. At present however, we shall treat empirical frequencies as probabilities and carry on the computations as above. The following table presents the relative frequencies of the elementary variables (3), the relative frequencies of the sum variables (5) and their v-measures (6), and the values of the components of covariation (7):

(1)	(2)				(3)	(4)	(5)	(6)	(7)
	$\underline{x}_1$	$\underline{x}_2$	$\underline{x}_3$	$\underline{x}_4$	p_m^4		q_m^4	α_m^4	γ_m^4
0	0	0	0	0	.248	0	0.0	0.0	0.0
1	1	0	0	0	.058	1	.467	2.360	.415
2	0	1	0	0	.022	2	.432	1.732	.010
3	1	1	0	0	.066	12	.227	.472	-.242
4	0	0	1	0	.058	3	.496	4.194	.596
5	1	0	1	0	.022	13	.395	1.356	-.019
6	0	1	1	0	.015	23	.372	1.184	-.100
7	1	1	1	0	.058	123	.503	4.444	.674
8	0	0	0	1	.029	4	.453	2.054	-.426
9	1	0	0	1	.015	14	.394	1.348	.087
10	0	1	0	1	.022	24	.343	1.066	.092
11	1	1	0	1	.044	124	.460	2.194	-.256
12	0	0	1	1	.102	34	.263	.648	.936
13	1	0	1	1	.036	134	.452	2.036	.067
14	0	1	1	1	.037	234	.417	1.560	-.080
15	1	1	1	1	.168	1234	.342	1.000	1.634

5.4. On the interpretation of the results.

We immediately see that some of the calculated components of covariation are negative. This contradicts the definition of γ's given in 5.1, and shows that the definition must be modified. This will be done in Section 6.

Considering only the absolute values of the components of covariation, it seems most natural to interpret the small components as representing "unnecessary dimensions" for the description of the data. Instead of the original 2^r-1 probabilities the distribution can be adequately described with fewer parameters replacing small components with zeros. To determine whether a component significantly differs from zero one should consider the sampling distribution of α^r.

At present we only observe that in our example the largest component corresponds to the covariation of all the variables. Fairly large are also the components of the covariation of $\underline{x}_3$ and $\underline{x}_4$ and of $\underline{x}_1$, $\underline{x}_2$, $\underline{x}_3$. It may be concluded that all the tests indeed measure some common dichotomous property, which, however, does not alone explain all the covariation.

5.5. Remarks.

Computationally the computing procedure is relatively light; except for additions and subtractions only multiplications by powers of 2 and a look-up table for logarithms is needed. The main difficulty is the rapid increase in the number of components when r increases. Applying the straightforward computation procedure it seems possible to handle at most 10-12 variables at the same time, which certainly is not enough in some applications. However, some modifications and refinements at the level of computational procedures seem feasible which can make possible the handling of more than a dozen variables.

6. The algebra of covariables.

As the starting point for the determination of the components of covariation we had the equation

$$\underline{a}^r = A_r \underline{z}^r \tag{1}$$

where $\underline{a}^r$ is the vector of sum variables of the given dichotomous variables $\underline{x}_1, \ldots, \underline{x}_r$, A_r the matrix defined on p. 9, and $\underline{z}^r$ a 2^r-dimensional vector of independent dichotomous variables. The question about the validity of the representation (1) was left unanswered above; we found out, however, that the derived equation

$$\alpha^r = A_r \gamma^r \tag{2}$$

for v-measures gives a solution γ^r whenever α^r is non-singular. Some of the components of γ^r may be negative, however, and this contradicts (1) and the definition of v-measure.

Mathematically there are presumably alternative ways to get out of the dilemma of negative γ's by suitably redefining either the $\underline{z}$-variables or the v-measure. One such way will be considered in 6.4. First, however, we deal with the case when all the components of γ^r are non-negative.

6.1. Measure algebra.

To illuminate the nature of the construction we use the concept of measure algebra.

Definition. Consider a finite Boolean algebra and a real-valued function μ, defined for the elements of B. If

(i) $-\infty < \mu \leq +\infty$

(ii) $\mu(x+y) = \mu(x) + \mu(y)$, when $xy = 0$, then μ is a measure and the ordered pair $[B,\mu]$ is a (finite) measure algebra.

If B is an infinite set, some continuity assumption for μ is needed to ensure the countable additivity of μ. It may be noted that we do not require that μ be non-negative.

As an example, consider the Boolean albegra $B(x_1,\ldots,x_r)$ of the dichotomous variables $\underline{x}_1,\ldots,\underline{x}_r$ and the probability measure p. From the properties of probability measure it follows that $[B(x_1,\ldots,x_r),p]$ is a measure algebra. The requirement (ii) implies that p is additive with respect to the variables whose product is the zero variable.

Consider now the sum variables $\underline{a}_m^r$, $m=0,\ldots,2^r-1$ of the given dichotomous variables $\underline{x}_1,\ldots,\underline{x}_r$. By theorem 4.1, v-measure is additive with respect to independent variables. In the following we are going to construct a Boolean algebra for which v fulfills the requirements of a measure. Addition will remain the same as before; as for multiplication, in this new algebra we require that the product of independent variables is zero.

Formally both structures, p-algebra and v-algebra, will be identical. The elementary variables in p-algebra correspond to independent covariables in v-algebra. Thus in p-algebra we have the decomposition of a variable into mutually exclusive elementary variables and in v-algebra into independent covariables.

6.3. The construction of v-algebra.

Now let 2^r arbitrary dichotomous variables $\underline{z}_o$, $\underline{z}_1,\ldots,\underline{z}_{2^r-1}$, completely independent of each other, be given. Consider the set of all linear combinations of the form

$$\underline{z} = \sum c_m \underline{z}_m \ , \quad c_m = 0 \text{ or } 1,$$

and denote this set by $B(z)$. Let $\underline{0}$ and $\underline{1}$ be the elements of $B(z)$ defined by

$$\underline{0} = 0\,\underline{z}_o + \cdots + 0\,\underline{z}_{2^r-1}$$

$$\underline{1} = 1\,\underline{z}_o + \cdots + 1\,\underline{z}_{2^r-1} \ .$$

If addition and multiplication for the elements of $B(z)$ are defined respectively by

$$\underline{z} + \underline{z}' = \sum \ (c_m + c_m')\underline{z}_m$$

$$\underline{z} * \underline{z}' = \sum \ (c_m \, c_m')\underline{z}_m$$

for any elements $\underline{z} = \Sigma\, c_m\, \underline{z}_m$, $z' = \Sigma\, c_m'\, \underline{z}_m'$, then it is easy to verify the requirements in 2.1 and show that $B(z)$ is a Boolean algebra.

Define also the v-measure for each element of $B(z)$ by

$$v(z) = \sum c_m v(z_m)$$

if $\underline{z} = \Sigma\, c_m\, \underline{z}_m$. Here the c_m's are taken as real numbers.

To show that $[B(z),v]$ is a measure algebra we have to show that the conditions (i), (ii) of 6.1 hold. Let $\underline{z} = \Sigma\, c_m\, \underline{z}_m$ and $\underline{z}' = \Sigma\, c_m'\, \underline{z}_m$. If $\underline{z} * \underline{z}' = 0$, then $c_m c_m' = 0$ for each m. Hence $v(z+z') = \Sigma\, (c_m+c_m')v(z_m) = \Sigma\, c_m v(z_m) + \Sigma\, c_m' v(z_m) = v(z) + v(z')$. Thus, because the condition (i) is obviously satisfied, $[B(z),v]$ is a measure algebra.

Returning now to the case of the given $\underline{x}_1,\ldots,\underline{x}_r$ and supposing that the variation vector α^r is non-singular, we can solve (2) for γ^r as indicated above. In this section we assume that all γ_m^r are non-negative.

Let now the independent variables $\underline{z}_o, \underline{z}_1, \ldots, \underline{z}_{2^r-1}$ be chosen so that

$$v(z_m) = \gamma_m^r \,, \quad m=0,\ldots,2^r-1 \tag{3}$$

and denote $\underline{z}^r = (\underline{z}_o, \ldots, \underline{z}_{2^r-1})$. Let Z be the set of all $\underline{z}^r$ with independent components for which (3) holds. For any given $\underline{z}^r$ we can construct the v-algebra $[B(z),v]$ as above, and it is easy to see that the v-algebras corresponding to different choices of $\underline{z}^r$ from Z are isomorphic.

Define the concept of v-equivalence as follows: the vectors $\underline{x}^r = (\underline{x}_o^r, \underline{x}_1^r, \ldots)$ and $\underline{y}^r = (\underline{y}_o^r, \underline{y}_1^r, \ldots)$ of dichotomous variables are v-equivalent, if $v(x_m^r) = v(y_m^r)$ for every m. This will be denoted by $\underline{x}^r \sim \underline{y}^r$. It follows now immediately from (2) that the vector $A_r \underline{z}^r$, whose components are elements of $B(z)$, is v-equivalent to the vector $\underline{a}^r$ of the sum variables of the original variables $\underline{x}_1, \ldots, \underline{x}_r$.

Instead of the original equation (1) we have thus established the relation of v-equivalence between $A_r \underline{z}^r$ and $\underline{a}^r$.

The results of the preceding considerations are expressed by the following proposition.

Proposition 6.1. Let $\underline{x}_1, \ldots, \underline{x}_r$ be dichotomous variables with non-singular variation vector α^r. Assume that all components of covariation are non-negative. Then it is possible to choose an ordered set of 2^r independent dichotomous variables $\underline{z}^r = (\underline{z}_o, \ldots, \underline{z}_{2^r-1})$, such that the vector $A_r \underline{z}^r$, whose components are elements of the v-algebra generated by $\underline{z}_o, \ldots, \underline{z}_{2^r-1}$, is v-equivalent to the vector $\underline{a}^r$ of the sum variables of the $\underline{x}_1, \ldots, \underline{x}_r$.

6.4. General case.

As before, suppose that the variation vector α^r of the variable $\underline{x}_1,\ldots,\underline{x}_r$ is non-singular. Consider now the case when some of the components of covariation may be negative.

The vector γ^r can be represented as a difference of two vectors with non-negative components, e.g., by defining

$$(4)\qquad \begin{aligned} &\phi^r_m = \gamma^r_m \quad\text{and}\quad \psi^r_m = 0\ , \quad\text{if}\quad \gamma^r_m \geq 0 \\ &\psi^r_m = -\gamma^r_m \quad\text{and}\quad \phi^r_m = 0\ , \quad\text{if}\quad \gamma^r_m < 0\ , \end{aligned}$$

$m=0,1,\ldots,2^r-1$. Denoting

$$\phi^r = (\phi^r_o,\ldots,\phi^r_{2^r-1})$$

$$\psi^r = (\psi^r_o,\ldots,\psi^r_{2^r-1})$$

we have $\gamma^r = \phi^r-\psi^r$.

The procedure of the preceding section can now be applied separately to ϕ^r and ψ^r because of the non-negativity of their components. Thus, we choose 2^{r+1} independent dichotomous variables $\underline{u}_m$, $\underline{w}_m$, $m=0,\ldots,2^r-1$, so that

$$v(u_m) = \phi^r_m\ , \quad v(w_m) = \psi^r_m\ , \quad m=0,\ldots,2^r-1$$

and construct the v-algebras $[B(u),v]$ and $[B(w),v]$.

Consider now the ordered pairs

$$\underline{z} = (\underline{u}, \underline{w})\ ,\ \underline{u} \in B(u)\ ,\ \underline{w} \in B(w)\ .$$

The set of these pairs will be denoted by $B(z)$. The zero and unit element in $B(z)$ are defined as $(\underline{0}, \underline{0})$ and $(\underline{1}, \underline{1})$, respectively, and addition and multiplication by

$$\begin{aligned} \underline{z} + \underline{z}' &= (\underline{u}, \underline{w}) + (\underline{u}', \underline{w}') \\ &= (\underline{u} + \underline{u}' , \underline{w} + \underline{w}') \\ \underline{z} * \underline{z}' &= (\underline{u} * \underline{u}' , \underline{w} * \underline{w}') \ . \end{aligned}$$

The properties of the operations on pairs thus defined remain the same as for the components, so that it is easy to verify the requirements of 2.1. Thus $B(z)$ is a Boolean algebra.

The v-measure for the elements of $B(z)$ is defined by

$$\begin{aligned} v(z) &= v((u,w)) \\ &= v(u)-v(w) \ . \end{aligned}$$

To show that $[B(z),v]$ is a measure algebra, form

$$\begin{aligned} v(z+z') &= v((u,w) + (u',w')) \\ &= v((u+u', w+w')) \\ &= v(u+u')-v(w+w') \ . \end{aligned}$$

Now $\underline{z} * \underline{z}' = \underline{0}$ implies that $\underline{u} * \underline{u}' = \underline{w} * \underline{w}' = \underline{0}$. Because $[B(u),v]$ and $[B(w),v]$ are measure algebras we further have

$$\begin{aligned} v(z+z') &= v(u)+v(u')-v(w)-v(w') \\ &= v((u,w))+v((u',w')) \\ &= v(z)+v(z') \ . \end{aligned}$$

Thus $[B(z),v]$ is a measure algebra (from the non-singularity of α^r it follows that all γ_m^r are finite and (1) is thus satisfied).

Form now the vector $\underline{z}^r$ of the pairs $(\underline{u}_m, \underline{w}_m)$, $m=0,\ldots,2^r-1$, and the vector $A_r\underline{z}^r$. We have

$$\begin{aligned} v(z_m^r) &= v(u_m)-v(w_m) \\ &= \phi_m^r - \psi_m^r \\ &= \gamma_m^r \end{aligned}$$

for each m, and from (2) it follows that $A_r\underline{z}^r$ is v-equivalent to the vector $\underline{a}^r$ of sum variables.

Summarizing the results we have the following proposition.

Proposition 6.2. Let $\underline{x}_1,\ldots,\underline{x}_r$ be dichotomous variables with non-singular variation vector α^r. It is then possible to choose an ordered set $\underline{z}^r$ of the pairs $(\underline{u}_m, \underline{w}_m)$, $m=0,\ldots,2^r-1$, of independent dichotomous variables, such that the vector $A_r\underline{z}^r$ is v-equivalent to the vector $\underline{a}^r$ of the sum variables of $\underline{x}_1,\ldots,\underline{x}_r$.

These results provide a mathematical framework, in terms of which it is possible to formulate, discuss, and evaluate various models, and assumptions about the joint distribution of the observed dichotomous variables $\underline{x}_1,\ldots,\underline{x}_r$.

REFERENCES

[1] McHugh, R.B., "Efficient estimation and local identification in latent class analysis," Psychometrica, Vol. 21 (1956), pp. 331-347.

[2] Peterson, W.W., Error-Correcting Codes, M.I.T. Press and Wiley, 1961.

[3] Solomon, H. (Editor), Studies in Item Analysis and Prediction, Stanford University Press, 1961.

ON SOME APPLICATIONS OF FINITE BOOLEAN ALGEBRAS TO DATA ANALYSIS

by

Juha Partanen

1. Introduction and Summary.

This preliminary report is concerned with the formal structure of two commonly used procedures in data analysis, analysis of variance and information or uncertainty analysis, described e.g., in [2], [11]-[15].

The formal similarity between the two procedures has been previously recognized ([3], [13]). The purpose of the present treatment is to show that they both may be characterized in terms of finite Boolean algebras.

For the analysis of variance this is achieved by introducing the concept of Boolean matrix algebra (section 3). This gives essentially a theory of idempotent matrices, whose importance to statistical analysis has been recently recognized (see e.g., [1], [5], [6]).

In section 4 it is indicated how this concept may be applied to analysis of variance. The approach gives a generalization of some earlier results ([4], [8]).

In section 5 a tentative definition is given to the concept Boolean p-algebra and this is applied to the analysis of contingency tables.

2. Boolean Algebra.

Consider a set $\mathfrak{S} = \{x,y,z,\dots\}$, which has two special elements, 0 and 1, and on which two binary operations called addition and multiplication are defined. This system is a Boolean algebra, if the following conditions are satisfied:

(i) $\mathfrak{B}$ is an abelian group with respect to addition, having 0 as the identity element.

(ii) Multiplication is associative and commutative with 1 as the identity element, and every element is idempotent, that is $1x = x$ and $xx = x$ for every $x \in \mathfrak{B}$.

(iii) Multiplication is distributive with respect to addition.

Two other useful concepts, complement and union, can be expressed in terms of addition and multiplication:

$$\bar{x} = 1 + x$$

$$x \cup y = x + y + xy$$

The axioms of Boolean algebra can be equivalently expressed e.g., in terms of multiplication, union and complement.

In the following we are concerned only with Boolean algebras having a finite number of elements.

A few more definitions are needed. A subset $(x_1,\ldots,x_r)$ of a Boolean algebra $\mathfrak{B}$ is called a set of generators if every element of $\mathfrak{B}$ can be expressed in terms of the $x_1,\ldots x_r$. The Boolean algebra generated by $x_1,\ldots,x_r$ is denoted by $\mathfrak{B}(x_1,\ldots,x_r)$.

We define also the basic elements of a Boolean algebra $\mathfrak{B}$: A non-zero element $b \in \mathfrak{B}$ is a basic element if $xb = b$ or 0 for all $x \in \mathfrak{B}$. From this definition it follows that $b_1b_2 = 0$ for two distinct basic elements b_1, b_2. In terms of the partial ordering "$x \leq y$" $\Longleftrightarrow$ $xy = x$ the basic elements may be characterized as the smallest non-zero elements of $\mathfrak{B}$.

Consider a Boolean algebra $\mathfrak{B}(x_1,\ldots,x_r)$. It can be seen that in this algebra the basic elements are products of r factors, each factor being either x_i or $\bar{x}_i, i=1,\ldots,r$. Clearly the number of basic elements can be at most 2^r. In the particular case when all 2^r products are non-zero, we say that the Boolean algebra $\mathfrak{B}(x_1,\ldots,x_r)$ is symmetric with respect to its generators $x_1,\ldots,x_r$.

Finally, we note that in a Boolean algebra $\mathfrak{B}$ every element can be uniquely expressed as a sum of the basic elements. For given an arbitrary $x \in \mathfrak{B}$, let $\Sigma' b_i$ be the sum of those basic elements b_i for which $xb_i = b_i$. We then have $x = \Sigma' b_i$ or $x + \Sigma' b_i = 0$, since $b_j(x + \Sigma' b_i) = 0$ for every basic element b_j.

3. Boolean matrix algebra.

Let a finite set $\mathfrak{B} = (A,B,C,\ldots)$ of $(v \times v)$ matrices with real elements be given. In general the matrices of $\mathfrak{B}$ need not be symmetric. Consider two operations: ordinary matrix multiplication and addition "$\oplus$", defined by

$$A \oplus B = A + B - 2AB.$$

The set $\mathfrak{B}$, together with these operations is called a Boolean matrix algebra if the following conditions hold:

(i) $\mathfrak{B}$ is closed with respect to addition and multiplication.

(ii) All elements of $\mathfrak{B}$ are idempotent and they commute with each other, that is $BB = B$ and $AB = BA$ for all $A, B \in \mathfrak{B}$.

(iii) There exists a unit matrix $U \in \mathfrak{B}$, so that $UB = B$ for every $B \in \mathfrak{B}$.

On the basis of these requirements one can verify the abstract conditions (i)-(iii) of the previous section, thus ensuring that the set $\mathfrak{B}$ together with the two operations is a Boolean algebra.

In particular it is seen that the role of zero element is taken by the $v \times v$ matrix having all its elements equal to zero. Also, from the existence of the unit matrix U it follows that U is unique. For if U, U_1 both are unit matrices we have $UU_1 = U = U_1$.

In accordance with the abstract notions we define matrix complement by $\bar{B} = U - B$ and matrix union by $A \cup B = A + B - AB$.

Example 1. Define the $v \times v$ matrix $\bar{J}_v$ by

$$\bar{J}_v = \frac{1}{v} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & & & \\ 1 & 1 & \cdots & 1 \end{bmatrix}$$

and let $J_v = I_v - \bar{J}_v$ (I_v is $v \times v$ identity matrix). It can be seen that the set $\{0_v, \bar{J}_v, J_v, I_v\}$ is a Boolean matrix algebra.

More generally, given an arbitrary set of "weights" $\{w_i, \Sigma w_i = 1, i = 1,\ldots,v\}$ define $\bar{W}$ by

$$\bar{W} = \begin{bmatrix} w_1 & w_2 & \cdots & w_v \\ w_1 & w_2 & \cdots & w_v \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ w_1 & w_2 & \cdots & w_v \end{bmatrix}$$

and $W = I_v - \bar{W}$. The matrices $0_v, \bar{W}, W, I_v$ also form a Boolean matrix algebra in which $\bar{W}W = W\bar{W} = 0$.

Example 2. Suppose all the matrices of a Boolean matrix algebra are diagonal. Because of idempotency the diagonal elements must be equal to either zero or one. We call this kind of Boolean matrix algebra a trivial Boolean matrix algebra.

3.1 Transformations.

Let a Boolean matrix algebra $\mathfrak{B}$ with the unit element U be given. There are various ways to define other Boolean matrix algebras related to $\mathfrak{B}$:

(i) The transposes $\{B', B \in \mathfrak{B}\}$ form a Boolean matrix algebra with the unit element U'.

(ii) Let $A \in \mathfrak{B}$. Then the set $\{AB, B \in \mathfrak{B}\}$ is a Boolean matrix algebra having A as the unit element.

(iii) Let $U = \Sigma\, C_m$ with $C_mC_n = 0$ for $m \neq n$. Then the sums of the matrices C_m form a Boolean matrix algebra with the matrices C_m as basic elements.

(iv) Let T be an arbitrary non-singular $v \times v$ matrix. Consider the set $\mathfrak{B}_T = \{T^{-1}BT, B \in \mathfrak{B}\}$ of matrices. This is a Boolean matrix algebra and the mapping $B \rightarrow T^{-1}BT$ shows that $\mathfrak{B}$ and $\mathfrak{B}_T$ are isomorphic. Following established terminology we say that $\mathfrak{B}$ and $\mathfrak{B}_T$ are similar. This relationship is obviously reflexive, symmetric and transitive.

In connection with the mapping $B \rightarrow T^{-1}BT$ we note the following theorem (for the proof of (i) below see e.g. Littlewood [10]):

Theorem 3.1. Let A_i, $i=1,\ldots,m$ be $v \times v$ matrices with

$$\begin{aligned} A_iA_j &= A_i \ , \ i = j \\ &= 0 \ , \ i \neq j \ . \end{aligned}$$

Then

(i) there exists a non-singular T such that for all A_i $T^{-1}A_iT$ is a diagonal matrix with all non-zero elements equal to one;

(ii) if both T and T_1 bring all A_i into diagonal form then there exists an orthogonal matrix C such that $T_1 = TC$;

(iii) if all A_i are symmetric, T can be taken to be an orthogonal matrix;

(iv) the rank of ΣA_i equals the sum of the ranks of the A_i.

This theorem can be applied to the basic elements $\{B_m\}$ of an arbitrary Boolean matrix algebra $\mathcal{B}$. Because of the isomorphism $B \to T^{-1}BT$ we have the corollary: Every Boolean matrix algebra is similar to a trivial Boolean matrix algebra.

3.2 Direct products.

We define the direct product of two matrices $B = [b_{ij}]$ and $C = [c_{ij}]$, with dimensions $m \times m$ and $n \times n$ respectively, as the $(mn \times mn)$ matrix

$$B \times C = \begin{bmatrix} c_{11}B & c_{12}B & \cdots & c_{1n}B \\ c_{21}B & c_{22}B & \cdots & \cdot \\ \vdots & & \cdots & \vdots \\ c_{n1}B & \cdot & \cdots & c_{nn}B \end{bmatrix}.$$

The following properties are direct consequences of the definition:

$$A \times (B \times C) = (A \times B) \times C \qquad \text{(associativity)}$$

$$A \times (B + C) = A \times B + A \times C \qquad \text{(distributive law)}$$

$$(A \times B)(C \times D) = (AC) \times (BD)$$

The last equality establishes a connection between direct products and ordinary products. In particular, if A^{-1}, B^{-1} exist, then $(A \times B)(A^{-1} \times B^{-1}) = (A^{-1} \times B^{-1})(A \times B) = I \times I$, so that $(A \times B)^{-1} = A^{-1} \times B^{-1}$.

Now let the Boolean matrix algebras $\mathcal{A}$ and $\mathcal{B}$ be given. Denote the basic elements of $\mathcal{A}$ and $\mathcal{B}$ by $\{A_m\}$, $\{B_n\}$, respectively. Consider the set of all sums of the form $\Sigma A_m \times B_n$. (It is assumed that in each sum all the terms are different from each other, and corresponding to the "empty" sum the zero matrix is taken.) This set of matrices can be easily shown to be a Boolean matrix algebra. It is called the direct product of $\mathcal{A}$ and $\mathcal{B}$, denoted by $\mathcal{A} \times \mathcal{B}$.

Example 3. Consider the Boolean matrix algebra $\mathcal{B}(W) = \{O_v, \overline{W}, W, I_v\}$ of the example 1, generated by the matrix W. Given two matrices W_1, W_2 of this type, with arbitrary dimensions, we consider the direct product $\mathcal{B}(W_1) \times \mathcal{B}(W_2)$. For this Boolean matrix algebra we use the notation $\mathcal{B}(W_1,W_2)$. The basic elements of $\mathcal{B}(W_1,W_2)$ are $\overline{W}_1 \times \overline{W}_2$, $W_1 \times \overline{W}_2$, $\overline{W}_1 \times W_2$, $W_1 \times W_2$.

The concept of direct product can be immediately generalized to an arbitrary finite number of factors. In particular, given the matrices $W_1,\ldots,W_r$ of the type of example 1, we may form the direct product $\mathcal{B}(W_1,\ldots,W_r) = \mathcal{B}(W_1) \times \cdots \times \mathcal{B}(W_r)$.

3.3 A definition of measure.

We recall that given a Boolean algebra $\mathcal{B}$ and a real-valued function μ, defined for all $x \in \mathcal{B}$, μ is called a measure if

(i) $-\infty < \mu(x) \leq +\infty$ for all $x \in \mathcal{B}$.

(ii) $\mu(x+y) = \mu(x) + \mu(y)$, if $xy = 0$.

This is a wide definition. Often it is also required that μ be finite and non-negative.

Consider now an arbitrary Boolean matrix algebra $\mathcal{B}$ of $v \times v$ matrices and an arbitrary v-dimensional vector y. Let T be a non-singular matrix of theorem 3.1 which brings all matrices of $\mathcal{B}$ into diagonal form. We wish to show that the squared norm

$$\|T^{-1}By\|^2 = (T^{-1}By)'(T^{-1}By)$$

satisfies the requirements of measure.

First of all we note that the proposed measure is not dependent on the particular T chosen to diagonalize $\mathfrak{B}$. For if $T_1 = TC$ also diagonalizes $\mathfrak{B}$ we have $\|T_1^{-1}By\| = \|C'T^{-1}By\| = \|T^{-1}By\|$ for all $B \in \mathfrak{B}$, y, since C is orthogonal.

It may also be seen that if all $B \in \mathfrak{B}$ are symmetric, the proposed measure simplifies to $\|T^{-1}By\|^2 = \|By\|^2$, since in this case T is an orthogonal matrix.

To prove the additivity (ii) let $A, B \in \mathfrak{B}$ and $AB = 0$. We have

$$\begin{aligned}\|T^{-1}(A \oplus B)y\|^2 &= \|T^{-1}(A+B)y\|^2 \\ &= y'(A+B)'T^{-1\prime}T^{-1}(A+B)y \\ &= \|T^{-1}Ay\|^2 + \|T^{-1}By\|^2 + y'A'T^{-1\prime}T^{-1}By \\ &\qquad + y'B'T^{-1\prime}T^{-1}Ay \,.\end{aligned}$$

The last two terms on the right are equal to zero. For $y'A'T^{-1\prime}T^{-1}By = y'T^{-1\prime}T'A'T^{-1\prime}T^{-1}BTT^{-1}y = (T^{-1}y)'(T^{-1}AT)'(T^{-1}BT)(T^{-1}y)$. Now $T^{-1}AT$ is diagonal, thus $(T^{-1}AT)'(T^{-1}BT) = (T^{-1}AT)(T^{-1}BT) = 0$. Similarly $y'B'T^{-1\prime}T^{-1}Ay = 0$.

Finally we note that the definition of measure could be slightly generalized by choosing an arbitrary diagonal matrix Q and considering the quadratic form $(T^{-1}By)'Q(T^{-1}By)$ instead of $(T^{-1}By)'(T^{-1}By)$. This modification in no way affects the above conclusions.

4. Application to decompositions of data.

We are now going to show that there exists a close connection between Boolean matrix algebras and certain commonly utilized decompositions of experimental data. To start with a very simple example, consider the Boolean matrix algebra $\mathfrak{B}(J) = \{O_v, J_v, \bar{J}_v, I_v\}$ of example 3.1. Let

$y = (y_1, \ldots, y_v)'$ be a v-dimensional vector. Then $\bar{J}_v y = (\bar{y}, \ldots, \bar{y})'$ and $J_v y = (y_1 - \bar{y}, \ldots, y_v - \bar{y})'$, where $\bar{y} = \frac{1}{v} \Sigma y_i$. The vector y is thus decomposed into two additive parts, the mean and the deviations of the mean, by applying the basic elements of $\mathfrak{B}(J)$ on y.

As another example, consider a partition of the components of y into two subsets, one containing the first r components and the other containing the rest. Let $\bar{\bar{y}}$, $\bar{y}_1$, $\bar{y}_2$ be the mean of all the components, the mean of those in the first set, and the mean of those in the second set, respectively. Consider the decomposition

$$y = \begin{bmatrix} \bar{\bar{y}} \\ \vdots \\ \bar{\bar{y}} \\ \bar{\bar{y}} \\ \vdots \\ \bar{\bar{y}} \end{bmatrix} + \begin{bmatrix} \bar{y}_1 - \bar{\bar{y}} \\ \vdots \quad \vdots \\ \bar{y}_1 - \bar{\bar{y}} \\ \bar{y}_2 - \bar{\bar{y}} \\ \vdots \quad \vdots \\ \bar{y}_2 - \bar{\bar{y}} \end{bmatrix} + \begin{bmatrix} y_1 - \bar{y}_1 \\ \vdots \quad \vdots \\ y_r - \bar{y}_1 \\ y_{r+1} - \bar{y}_2 \\ \vdots \quad \vdots \\ y_v - \bar{y}_2 \end{bmatrix}$$

or in matrix notation

$$y = B_o y + B_1 y + B_2 y \ ,$$

where $B_o = \bar{J}_v$

$$B_1 = \begin{bmatrix} \bar{J}_r & 0 \\ 0 & \bar{J}_{v-r} \end{bmatrix} - \bar{J}_v$$

$$B_2 = \begin{bmatrix} J_r & 0 \\ 0 & J_{v-r} \end{bmatrix} .$$

It is easy to verify that the matrices B_o, B_1, B_2 are idempotent and $B_i B_j = 0$ with $i \neq j$; they can thus be taken as basic elements of a Boolean matrix algebra.

Obviously a similar argument applies to more general nested or hierarchical designs. Some generalization can be achieved in another direction, too: the matrices $\overline{J}$, which correspond to simple averages, can be replaced by matrices $\overline{W}$, formed by using an arbitrary set of weights $\{w_i, \Sigma\, w_i = 1\}$.

We turn now to the consideration of factorial decomposition, by which we mean the familiar decomposition of observations into over-all mean, main effects and interactions. Consider a factorial experiment with r factors $f_1,\ldots,f_r$, and let n_i be the number of levels for f_i. The number of treatment combinations is then $v = \pi\, n_i$. We denote the treatment combinations, ordered lexicographically, by the index $j = (j_1,\ldots,j_r)$, $j_i = 1,\ldots,n_i$, $j = 1,\ldots,v$. The unknown treatment effects on the observed variable are denoted by τ_j, $j=1,\ldots,v$. Let $\tau = (\tau_1,\ldots,\tau_v)'$. The purpose of an experiment is to obtain an estimate of τ and use it to provide information about the magnitudes of the various main effects and interactions. We are here concerned only with the "fixed effects" models.

As we are going to show using the 2-factor case as an example, the decomposition of τ into over-all mean, main effects and interactions can be obtained by applying the basic elements of the Boolean matrix algebra $\mathfrak{B}(J_1,\ldots,J_r)$, as defined in 3.2, to the vector τ.

Consider first the 2-factor case. The basic elements of $\mathfrak{B}(J_1,J_2)$ are $B_o = \overline{J}_1 \times \overline{J}_2$, $B_1 = J_1 \times \overline{J}_2$, $B_2 = \overline{J}_1 \times J_2$, $B_3 = J_1 \times J_2$. The notations may be brought closer to those usually employed when dealing with symmetric Boolean algebras by putting $F_1 = J_1 \times I$, $F_2 = I \times J_2$. The basic elements can be then expressed as $B_o = \overline{F}_1\overline{F}_2$, $B_1 = F_1\overline{F}_2$, $B_2 = \overline{F}_1 F_2$, $B_3 = F_1F_2$.

Denoting now $\tau_{..} = (1/n_1n_2)\Sigma\,\tau_{ij}$, $\tau_{i.} = (1/n_2)\Sigma_j\tau_{ij}$, $\tau_{.j} = (1/n_1)\Sigma_i\tau_{ij}$, it can be verified that

$$\bar{F}_1\bar{F}_2\tau = (\tau_{..})$$

$$F_1\bar{F}_2\tau = (\tau_{i.}-\tau_{..})$$

$$\bar{F}_1F_2\tau = (\tau_{.j}-\tau_{..})$$

$$F_1F_2\tau = (\tau_{ij}-\tau_{i.}-\tau_{.j}+\tau_{..})\ .$$

It should be noted that all the vectors on the right-hand side are n_1n_2-dimensional. Thus e.g., $(\tau_{i.}-\tau_{..})$ denotes the vector having $\tau_{i.}-\tau_{..}$ in its $(i,j)^{th}$ position for every $j = 1,\dots,n_2$. This formulation causes some repetition; it may be noted, however, that when actual computations are performed one can replace the matrix $\bar{J}_1$ by one of its identical rows, and thus reduce the dimension.

It is of some interest to observe the form of $B\tau$ for some other matrices $B \in \mathcal{B}(J_1,J_2)$. We have

$$F_1\tau = (\tau_{ij}-\tau_{.j})$$

$$F_2\tau = (\tau_{ij}-\tau_{i.})$$

$$(F_1 \oplus F_2)\tau = ((\tau_{i.}-\tau_{..})+(\tau_{.j}-\tau_{..}))$$

$$(F_1 \cup F_2)\tau = (\tau_{ij}-\tau_{..})\ .$$

In the general case we have r factors $f_1,\dots,f_r$. Consider the Boolean matrix algebra $\mathcal{B}(J_1,\dots,J_r)$. As above, we put $F_1 = J_1 \times I \times \cdots \times I$, $F_2 = I \times J_2 \times \cdots \times I, \dots, F_r = I \times I \times \cdots \times J_r$. The basic elements of $\mathcal{B}(J_1,\dots,J_r)$ can then be written $B_0 = \bar{F}_1\bar{F}_2\dots\bar{F}_r$, $B_1 = F_1\bar{F}_2\cdots\bar{F}_r,\dots,$ $B_{2^r-1} = F_1F_2\dots F_r$, and we have, with obvious notations,

$$\bar{F}_1\bar{F}_2\ldots\bar{F}_r\tau = (\tau_{\cdot\cdot\cdot})$$

$$F_1\bar{F}_2\ldots\bar{F}_r\tau = (\tau_{i_1\cdot\cdot}-\tau_{\cdot\cdot\cdot})$$

$$\bar{F}_1F_2\ldots\bar{F}_r\tau = (\tau_{\cdot i_2\cdot}-\tau_{\cdot\cdot\cdot})$$

$$F_1F_2\ldots\bar{F}_r\tau = (\tau_{i_1i_2\cdot}-\tau_{i_1\cdot\cdot}-\tau_{\cdot i_2\cdot}+\tau_{\cdot\cdot\cdot})$$

. . . .

$$F_1F_2\ldots F_r\tau = ((r-1)^{\text{th}} \text{ order interactions}).$$

Formally the same results are obtained if the matrices $J_1,\ldots,J_r$ are replaced by the matrices $W_1,\ldots,W_r$, given by arbitrary sets of weights $\{w^1_{ij}, \Sigma w^1_{ij} = 1\},\ldots,\{w^r_{ir}, \Sigma w^r_{ir} = 1\}$. The τ-averages, in this case, are weighted averages.

Suppose now that using one of the many available experimental designs the vector τ of treatment effects is estimated by a vector t. It is usually assumed that t is normally distributed with the mean $Et = \tau$. We denote the covariance matrix of t by $V(t)$. $V(t)$ depends both on the error variance and the experimental design. In the simplest case, when all the treatments are in one block, with equal number of observations per treatment combination, say k, and a uniform error variance σ^2, we have $V(t) = (\sigma^2/k)I$.

For any $B \in \mathfrak{B}(J_1,\ldots,J_r)$, we take Bt as an estimate of $B\tau$. Bt is normally distributed with the mean $EBt = B\tau$ and the covariance matrix $V(Bt) = BV(t)B'$. In the simple case $V(t) = (\sigma^2/k)I$ this reduces to $V(Bt) = (\sigma^2/k)B$. We also have $\mathrm{Cov}(Bt,Ct) = BV(t)C' = (\sigma^2/k)BC$ for arbitrary $B,C \in \mathfrak{B}(J_1,\ldots,J_r)$. In particular we see that Bt and Ct are independent whenever $BC = 0$.

These results also give the decomposition of the total sum of squares. For since $I = \Sigma B_m$, $B_m B_n = 0$, we have $\|t\|^2 = \Sigma \|B_m t\|^2$. Cochran's theorem implies that when $\|t\|^2$ is distributed as $\sigma^2 \chi^2$ with v degrees of freedom and the non-centrality parameter $\|\tau\|$, then $\|B_m t\|^2$ is distributed as $\sigma^2 \chi^2$ with rank (B_m) degrees of freedom and the non-centrality parameter $\|B_m \tau\|$.

More generally it may be seen that $V(Bt)$, $B \in \mathfrak{B}(J_1,\ldots,J_r)$ assumes a computationally simple form if $V(t)$ can be written in the form

$$V(t) = \Sigma \lambda_m B_m$$

for some constants $\{\lambda_m\}$, $m = 0,\ldots,2^r-1$. This condition, equivalent to the condition obtained in [8], has been shown to hold for many commonly used experimental designs ([8], [9]).

5. Boolean algebras on contingency tables.

In this section our concern is with random experiments in which the observed variables are random variables in the widest sense, possessing no metric structure. The common way to describe the structure of the experiment is in the form of a contingency table.

Let r be the number of observed variables, each with a finite set of possible values labelled j_i, $j_i = 1,\ldots,n_i$, $i=1,\ldots,r$. The number of different value combinations or elementary events is $v = \pi n_i$. These will be ordered lexicographically and denoted by $j = (j_1,\ldots,j_r)$, $j=1,\ldots,v$. The probability structure of the experiment is determined by the vector of probabilities of the elementary events, denoted by $p = (p_1,\ldots,p_v)'$.

To describe various aspects of the total distribution p, especially dependencies between the observed variables, various other vectors which are functions of p may be obtained. Here we consider

(i) marginal distributions, obtained by summing over all values of some of the variables;

(ii) conditional distributions; and

(iii) interactions. This term has been given various non-equivalent meanings; in the present context we mean by interactions v-dimensional vectors which have the form $(p_{ij}/p_{i.}p_{.j})$ in the 2-variable case, $(p_{ijk}p_{i..}p_{.j.}p_{..k}/p_{ij.}p_{i.k}p_{.jk})$ in the 3-variable case etc. Given a distribution of more than two variables, interactions may be formed using both the marginal distributions and the conditional distributions.

To achieve symmetry all the vectors mentioned above are treated as v-dimensional vectors. In the case of marginal distribution this means that e.g., the vector $(p_{j_1 \cdots})$ has $p_{j_1 \cdots}$ as its $(j_1, j_2, \ldots, j_r)^{th}$ component for all $j_2, \ldots, j_r$.

Our purpose is to show that this somewhat loosely characterized collection of vectors, all derived from the distribution p, may be treated as a Boolean algebra. In addition, the Shannon-Wiener information measure may be defined for the elements of this Boolean algebra and it satisfies the general requirements of measure.

5.1 Boolean p-algebra.

We give the following

Definition. Let $\mathfrak{B}$ be a finite Boolean algebra. Suppose that with each element $x \in B$ is associated a v-dimensional vector $p(x) = (p_1(x), \ldots, p_v(x))'$, with non-negative components. The set $\{p(x), x \in \mathfrak{B}\}$

is called a Boolean p-algebra if

$$p_j(x+y) = p_j(x)p_j(y) \quad \text{for all} \quad j=1,\ldots,v, \text{ whenever } xy = 0.$$

From this definition it follows immediately that $p_j(o) = 1$ for all $j=1,\ldots,v$ (assuming that not all $p(x) = 0$).

To find concrete examples of Boolean p-algebras we first take up the case of two variables. Among the vectors derived from the total distribution $p = (p_{ij})$ we have the marginal distributions $(p_{i.})$ and $(p_{.j})$, the conditional distributions $(p_{ij}/p_{.j})$ and $(p_{ij}/p_{i.})$, and the interactions $(p_{ij}/p_{i.}p_{.j})$.

Now let $\mathfrak{B}(x_1,x_2)$ be the symmetric Boolean algebra generated by x_1 and x_2. To each element $x \in \mathfrak{B}(x_1,x_2)$ we assign a vector $p(x)$ by setting

$$p(\bar{x}_1\bar{x}_2) = (1)$$

$$p(x_1\bar{x}_2) = (p_{i.})$$

$$p(\bar{x}_1x_2) = (p_{.j})$$

$$p(x_1x_2) = (p_{ij}/p_{i.}p_{.j})$$

for the basic elements of $\mathfrak{B}(x_1,x_2)$ and using the additivity condition to determine the vectors associated with other elements of $\mathfrak{B}(x_1x_2)$. Since every element in a Boolean algebra can be uniquely expressed as a sum of basic elements, this assignment is unique for all $x \in \mathfrak{B}(x_1,x_2)$. We have in particular

$$p(x_1) = (p_{ij}/p_{.j})$$
$$p(x_2) = (p_{ij}/p_{i.})$$
$$p(x_1+x_2) = (p_{i.}p_{.j})$$
$$p(x_1 \cup x_2) = (p_{ij})$$
$$p(1) = (p_{ij}) \; .$$

We turn now to the case of three variables. Using obvious notations we assign the following vectors to the basic elements of $\mathfrak{B}(x_1,x_2,x_3)$:

$$p(\bar{x}_1\bar{x}_2\bar{x}_3) = (1)$$
$$p(x_1\bar{x}_2\bar{x}_3) = (p_{i..})$$
$$p(\bar{x}_1x_2\bar{x}_3) = (p_{.j.})$$
$$p(x_1x_2\bar{x}_3) = (p_{ij.}/p_{i..}p_{.j.})$$
$$p(\bar{x}_1\bar{x}_2x_3) = (p_{..k})$$
$$p(x_1\bar{x}_2x_3) = (p_{i.k}/p_{i..}p_{..k})$$
$$p(\bar{x}_1x_2x_3) = (p_{.jk}/p_{.j.}p_{..k})$$
$$p(x_1x_2x_3) = (p_{ijk}p_{i..}p_{.j.}p_{..k}/p_{ij.}p_{i.k}p_{.jk}) \; .$$

For other elements we then have expressions like

$$p(x_1) = (p_{ijk}/p_{.jk})$$
$$p(x_1x_2) = (p_{..k}p_{ijk}/p_{i.k}p_{.jk})$$
$$p(x_1+x_2) = ((p_{i.k}/p_{..k})(p_{.jk}/p_{..k}))$$
$$p(x_1 \cup x_2) = (p_{ijk}/p_{..k})$$
$$p(x_1 \cup x_2 \cup x_3) = (p_{ijk})$$
$$p(1) = (p_{ijk}) \; ,$$

which all have obvious probabilistic meanings. The general rule is thus to assign marginal distributions and interactions to basic elements.

The conditional distributions and interactions then correspond to other elements of the symmetric Boolean algebra.

5.2 A definition of measure.

Formally the treatment just presented may be interpreted as a calculus for probability vectors derived from p. In some applications it is possible to consider the probability vectors as describing various identifiable "sources of variation". In those cases a summary description of sources of variation and "strengths of association" between variables in terms of a real valued measure may be helpful.

Consider the quantity

$$H(x) = -\Sigma_j \; p_j(1) \ln p_j(x)$$

defined for every element $p(x)$ of a Boolean p-algebra. From the additivity condition of 5.1 it follows that $H(x+y) = H(x) + H(y)$ whenever $xy = 0$.

In the case of two variables we have

$$H(\bar{x}_1\bar{x}_2) = 0$$

$$H(x_1\bar{x}_2) = -\Sigma \; p_{i.} \ln p_{i.}$$

$$H(\bar{x}_1 x_2) = -\Sigma \; p_{.j} \ln p_{.j}$$

$$H(x_1 x_2) = -\Sigma \; p_{ij} \ln (p_{ij}/p_{i.}p_{.j}) \; .$$

Thus $H(x_1\bar{x}_2)$ and $H(\bar{x}_1 x_2)$ give the Shannon-Wiener information measures of the marginal distributions. The quantity $-H(x_1x_2)$ (which is always non-negative) is sometimes called redundancy.

Similar expressions may be obtained in more general cases. The relationships between all these information or uncertainty components have been extensively treated (e.g., [11]-[13], [15]). It may thus be of some interest to observe that all the relationships can be accounted for by the fact that H is a measure on a finite Boolean algebra.

5.3 A generalization.

The above considerations can be slightly generalized. It may be noted that we obtained a multiplicative decomposition of the total distribution p. Now let q be another v-dimensional vector, which may be thought of as a distribution specified by an alternative hypothesis. In this case it is reasonable to modify the procedure so that it will give a decomposition of the deviation between p and q, described e.g., by the vector of ratios $(p_1/q_1,\ldots,p_v/q_v)$. (In the analysis of variance the same problem does not exist, since the linear decomposition determined by a Boolean matrix algebra may be directly applied to the difference between the two vectors.)

In general we note that if $\{p(x),\ x \in \mathfrak{B}\}$, $\{q(x),\ x \in \mathfrak{B}\}$ are two Boolean p-algebras of v-dimensional vectors, both determined by the same abstract Boolean algebra $\mathfrak{B}$, then the vectors $\{p(x)^{-1} = (p_1(x)^{-1},\ldots, p_v(x)^{-1}),\ x \in \mathfrak{B}\}$ and $\{p(x)*q(x) = (p_1(x)q_1(x),\ldots,p_v(x)q_v(x)),\ x \in \mathfrak{B}\}$ form Boolean p-algebras. Then also $\{p(x)*q(x)^{-1} = (p_1(x)/q_1(x),\ldots, p_v(x)/q_v(x)),\ x \in \mathfrak{B}\}$ is a Boolean p-algebra.

We consider some specific examples in the case of three variables. In the first instance let q be the joint distribution of three independent variables having as marginal distributions those determined by p. In the second take q as the uniform distribution. The expressions for the basic elements of $\{p(x)*q(x)^{-1}\}$ in these cases become

	(i) (independence)	(ii) (uniform)
$\bar{x}_1\bar{x}_2\bar{x}_3$	(1)	(1)
$x_1\bar{x}_2\bar{x}_3$	(1)	$(n_1 p_{i..})$
$\bar{x}_1 x_2\bar{x}_3$	(1)	$(n_2 p_{.j.})$
$x_1 x_2\bar{x}_3$	$(p_{ij.}/p_{i..}p_{.j.})$	$(p_{ij.}/p_{i..}p_{.j.})$
$\bar{x}_1\bar{x}_2 x_3$	(1)	$(n_3 p_{..k})$
$x_1\bar{x}_2 x_3$	$(p_{i.k}/p_{i..}p_{..k})$	$(p_{i.k}/p_{i..}p_{..k})$
$\bar{x}_1 x_2 x_3$	$(p_{.jk}/p_{.j.}p_{..k})$	$(p_{.jk}/p_{.j.}p_{..k})$
$x_1 x_2 x_3$	$(p_{ijk}p_{i..}p_{.j.}p_{..k}/p_{ij.}p_{i.k}p_{.jk})$	$(p_{ijk}p_{i..}p_{.j.}p_{..k}/p_{ij.}p_{i.k}p_{.jk})$

Under (i) the product of all the basic elements is $(p_{ijk}/p_{i..}p_{.j.}p_{..k})$. This describes the total deviation from independence in the contingency table. Under (ii), the total deviation of p from the uniform distribution is decomposed into components three of which describe the difference in marginal distributions, the rest being due to non-independence.

It is also possible to define H-measure on Boolean p-algebras of the type $\{p(x)*q(x)^{-1}, x \in \mathfrak{B}\}$ by

$$H(x) = -\Sigma\, p_j(1)\ln(p_j(x)/q_j(x)).$$

The analysis of these quantities has been carried through by Kullback [7] for several special cases.

REFERENCES

[1] Banerjee, K.S. (1964), "A Note On Idempotent Matrices," Ann. Math. Stat., 35, 880-882

[2] Garner, W.R. (1962), Uncertainty and Structure as Psychological Concepts, John Wiley and Sons

[3] Garner, W.R. and McGill, W.J. (1956), "Relation Between Information and Variance Analyses," Psychometrika, 21, 219-228

[4] Good, I.J. (1958), "The Interaction Algorithm and Practical Fourier Analysis," Jour. Royal Stat. Soc., 20, 361-372

[5] Graybill, F.A. and Marsaglia, G. (1957), "Idempotent Matrices and Quadratic Forms in the General Linear Hypothesis," Ann. Math. Stat., 28, 678-686

[6] Hogg, R.V., and Craig, A.T. (1958), "On the Decomposition of Certain Variables," Ann. Math. Stat., 29, 608-610

[7] Kullback, S. (1959), Information Theory and Statistics, John Wiley and Sons

[8] Kurkjian, B., and Zelen, M. (1962), "A Calculus for Factorial Arrangements," Ann. Math. Stat., 33, 600-619

[9] Kurkjian, B., and Zelen, M. (1963), "Applicationsof the Calculus of Factorial Arrangements I," Biometrica, 50, 63-73

[10] Littlewood, D.E. (1958), The Theory of Group Characters, Oxford Univ. Press, 2nd edition

[11] Luce, R.D. (ed.) (1960), Developments in Mathematical Psychology, Free Press.

[12] McGill, W.J. (1954), "Multivariate Information Transmission," Psychometrika, 19, 97-116

[13] Quastler, H. (ed.) (1955), Information Theory in Psychology, Free Press

[14] Toda, M.(1963), "Review of Garner," [2] Psychometrika, 28, 293-310

[15] Watanabe, S., (1960), "Information Theoretical Analysis of Multivariate Correlation," IBM Journ. of Res. and Devel. 4, 66-82